UNIVERSITY OF CALIFORNIA
RIVERSIDE

Learning to Attack, Protect, and Enhance Deep Networks

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Zikui Cai

June 2024

Dissertation Committee:

Dr. M. Salman Asif, Chairperson
Dr. Amit K. Roy-Chowdhury
Dr. Konstantinos Karydis

The Dissertation of Zikui Cai is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor, Dr. Salman Asif, whose unwavering support and mentorship have been instrumental throughout my doctoral journey. His guidance in research and his influence on my thinking have been invaluable.

I would like to extend my sincere appreciation to my other mentors and collaborators, across academia and industry, including Dr. Amit K Roy-Chowdhury, Dr. Chengyu Song, Dr. Srikanth V. Krishnamurthy, Dr. Shantanu Rane, Dr. Alejandro E. Brito, Dr. Ziyan Wu, Dr. Benjamin Planche, Dr. Zhongpai Gao, Dr. Meng Zheng, and Dr. Terrence Chen. I am also grateful to my PhD committee members, Dr. Amit K Roy-Chowdhury and Dr. Konstantinos Karydis, for their invaluable contributions and insights during my dissertation.

My journey has been greatly enriched by the camaraderie, support, and kindness of my friends at UCR and beyond. I am also thankful to UCR Outdoor Excursions for the wonderful trips to the mountains and the ocean, which provided much-needed respite, adventure, and rejuvenation. Furthermore, I would like to extend my appreciation to the Tennis Club, Aikido Dojo, and International Student Union for providing vital recreational activities and a sense of community.

Thank you all for being part of this incredible journey!

To my parents for all the love and support.

ABSTRACT OF THE DISSERTATION

Learning to Attack, Protect, and Enhance Deep Networks

by

Zikui Cai

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, June 2024
Dr. M. Salman Asif, Chairperson

Artificial intelligence (AI) systems have demonstrated remarkable capabilities, yet concerns about their security and safe deployment persist. With the rapid adoption of AI across critical domains, ensuring the robustness and reliability of these models is imperative. This research addresses this challenge by exposing vulnerabilities in AI systems and enhancing their trustworthiness. By systematically uncovering flaws, it aims to raise awareness of the precautions necessary for utilizing AI in high-stakes scenarios. The methodology involves identifying vulnerabilities, quantifying worst-case performance via attacks, and generalizing insights to practical deployment settings. Additionally, it investigates techniques to strengthen model trustworthiness in real-world scenarios, contributing to rigorous AI safety research that promotes responsible and beneficial system development. Specifically, this research reveals vulnerabilities in neural networks by developing efficient black-box attacks on various deep learning models across different tasks. Additionally, it focuses on improving AI trustworthiness by detecting adversarial examples using language models and enhancing user privacy through innovative facial de-identification methods.

For highly effective black-box attacks, ensemble-based and context-aware approaches were developed. These methods optimize over ensemble model weight spaces to craft adversarial examples with extreme efficiency, significantly outperforming existing input space attacks. Multi-modal testing demonstrated that these attacks could fool systems on diverse tasks, highlighting the need to evaluate deployment robustness against such methods. Additionally, by weaponizing context to manipulate statistical relationships that models rely on, context-aware attacks were shown to profoundly mislead systems, revealing reasoning vulnerabilities.

To protect user privacy, an algorithm was developed for seamlessly de-identifying facial images while retaining utility for downstream tasks. This approach, grounded in differential privacy and ensemble learning, maximizes obfuscation and non-invertibility to prevent re-identification. By disentangling identity attributes from utility attributes like expressions, the method significantly enhances de-identification rates while preserving utility.

To enhance the robustness and efficiency of computational imaging pipelines, including Fourier phase retrieval and coded diffraction imaging, I developed a framework that learns reference signals or illumination patterns using a small number of training images. This framework employs an unrolled network as a solver. Once learned, the reference signals or illumination patterns serve as priors, significantly improving the efficiency of signal reconstruction.

Overall, this research contributes to a more secure and reliable deployment of AI systems, ensuring their safe and beneficial use across critical domains.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial intelligence (AI) systems have achieved remarkable capabilities, yet concerns about their security and safety persist. Neural networks have been shown vulnerable to adversarial attacks that lead to arbitrarily wrong outputs with slight manipulations to inputs [438, 165, 326, 66, 64]. As AI adoption rapidly increases across critical domains, ensuring model robustness and reliability is imperative. I will present some of my efforts aiming to address this pressing challenge along with how such problems are previously approached.

1.1 Attacking deep networks: harnessing surrogates and weaponizing context

Adversarial attacks on deep learning models have emerged as a significant area of research, highlighting vulnerabilities in these systems and underscoring the importance of robust model development. This section discusses two sophisticated approaches that leverage surrogate-

based strategies to execute effective attacks on deep networks, focusing on image classifiers and object detection systems.

Blackbox Attacks via Surrogate Ensemble Search (BASES). The first approach, known as Blackbox Attacks via Surrogate Ensemble Search (BASES), addresses the challenge of performing efficient and successful adversarial attacks without extensive interaction with the target model. Traditional blackbox methods often suffer from low success rates and high query demands, particularly in targeted attacks. BASES innovates by utilizing a compact set of surrogate models to approximate the decision boundaries of the target model, thereby allowing for the generation of adversarial examples with significantly fewer queries.

The core of the BASES method involves a perturbation machine, which generates perturbed images by minimizing a weighted loss function. This function is dynamically adjusted over the surrogate models based on the feedback from a limited number of queries to the victim model. Crucially, the dimensionality of the search space in BASES corresponds directly to the number of surrogate models used, which drastically reduces the number of required queries. Empirical results demonstrate that BASES can achieve a success rate exceeding 90% in targeted attacks with as few as three queries per image on average, and an impressive 99% success rate for untargeted attacks with only 1-2 queries per image. This method not only outperforms existing blackbox strategies in terms of efficiency but also maintains high transferability of the perturbations, making it suitable for hard-label blackbox attacks.

Context-Aware Adversarial Attacks for Object Detection. The second approach extends the concept of surrogate-based strategies to the realm of object detection, where adversarial attacks must consider the contextual relationships within an image. Object detectors analyze multi-

ple aspects of an image, such as the presence and interactions of various objects, making them more complex targets for adversarial attacks than standard image classifiers.

This method utilizes the co-occurrence of objects along with their relative locations and sizes to inform the generation of adversarial examples. By incorporating these contextual details into the attack generation process, it becomes possible to manipulate object detectors into misclassifying or failing to detect certain objects, thereby achieving targeted mis-categorization. This approach has shown to improve the transfer success rates of attacks on blackbox object detectors significantly. When tested on popular datasets such as PASCAL VOC and MS COCO, the technique demonstrated up to a 20 percentage point improvement over existing state-of-the-art methods.

Both strategies underscore the potential of surrogate-based methods in crafting more effective adversarial attacks that can adapt to the complexities and constraints of different deep learning architectures. By reducing reliance on extensive query mechanisms and incorporating a deeper understanding of model behavior and image context, these approaches offer a more nuanced and powerful toolkit for the adversarial testing of deep networks.

1.2 Protecting deep networks: advancing privacy with disguise

As deep learning continues to permeate various sectors, including those involving sensitive information, the imperative to protect privacy in AI-driven systems has become paramount. This section introduces ”Disguise,” a novel algorithm designed to shield individuals' identities in images while preserving the utility of these data for deep learning tasks. The approach is grounded in principles of differential privacy and ensemble learning, representing a significant step forward in the ethical use of AI technologies.

Concept and Motivation. Traditional de-identification methods often struggle with the dual challenge of maintaining privacy while preserving the analytical usability of data. Common techniques, such as pixelation or masking, tend to degrade the quality of the data, which can detrimentally affect the performance of AI models that rely on detailed visual inputs. Disguise addresses these issues by employing a sophisticated algorithm that can selectively alter identifying features in images, thereby ensuring privacy without compromising the data's utility for subsequent analysis.

Algorithmic Framework. Disguise operates through a combination of variational techniques and a mixture-of-experts model. The core mechanism involves extracting identifying features from facial images and substituting them with synthetic, non-reversible alternatives generated through variational methods. This substitution is designed to maximize the obfuscation of personal identities, ensuring that the modified images are resistant to reverse engineering and re-identification.

To preserve the usefulness of de-identified images for deep learning applications, Disguise leverages a mixture-of-experts that supervises the modification process. This supervision helps maintain important non-identifying attributes such as age, gender, and emotion, which are crucial for tasks like demographic analysis and sentiment detection. Each expert in the ensemble focuses on a specific attribute, ensuring that the essential characteristics of the data are retained even after identity obfuscation.

Evaluation and Impact. Extensive evaluations of Disguise have demonstrated its effectiveness across multiple datasets. The algorithm not only achieves higher de-identification rates but also ensures superior consistency in preserving non-identifying attributes compared to previous approaches. These results highlight Disguise's potential to facilitate the broader adoption of

privacy-preserving technologies in AI, particularly in fields like healthcare and social media where user consent and data sensitivity are critical.

Furthermore, Disguise's implementation of differential privacy principles enhances its appeal, as it provides a quantifiable measure of privacy that complies with emerging regulations and standards in data protection. This makes it an invaluable tool for organizations seeking to deploy AI solutions that require robust privacy guarantees without sacrificing data utility.

Overall, the Disguise algorithm represents a significant advancement in the field of AI privacy, offering a practical solution that carefully balances the needs for both robust data protection and high-quality analytical outputs in deep learning systems. Its development not only addresses critical ethical concerns but also opens up new possibilities for the responsible use of AI in sensitive domains.

1.3 Enhancing deep networks: data driven signal processing

The enhancement of deep networks through sophisticated signal processing techniques represents a pivotal advancement in improving the performance and applicability of artificial intelligence systems. This section delves into the cutting-edge methodologies that leverage data-driven approaches to optimize signal processing in two main areas: Fourier phase retrieval and coded diffraction pattern recovery. These enhancements not only improve the accuracy and efficiency of deep learning models but also broaden their application scope across various scientific and industrial domains.

Fourier Phase Retrieval with Learned References Fourier phase retrieval is a fundamental problem in imaging and optics, where the goal is to reconstruct a signal from the magnitudes

of its Fourier transform. Traditional approaches to this problem often rely on iterative algorithms that can suffer from slow convergence and sensitivity to initialization. Our approach revolutionizes this process by introducing learned references, which are integrated into the Fourier measurement process to guide and accelerate convergence.

By embedding a learned reference in the Fourier amplitude measurements, our method effectively constrains the solution space, reducing ambiguities and enhancing the recovery accuracy. The reference signal is optimized through a data-driven process using backpropagation, adapting to specific characteristics of the target data set. This method significantly simplifies the phase retrieval process by reducing the number of required iterations, thus lowering computational costs while maintaining high reconstruction fidelity.

Data-driven Illumination Patterns for Coded Diffraction Imaging. Expanding the scope of signal processing to more complex scenarios, our work on signal recovery from nonlinear measurements addresses the challenges posed by coded diffraction patterns. In practical applications, such as X-ray crystallography or diffraction imaging, the signal is modulated by a sequence of codes before sensor measurement, complicating the recovery process.

Our novel framework optimizes the sensing parameters, particularly the illumination patterns, to significantly improve the quality of the recovered signal. By representing the phase retrieval process as an unrolled network with a fixed number of layers, we can directly optimize the measurement parameters to minimize recovery errors. This approach ensures that each iteration, or layer, contributes optimally to the recovery process, making efficient use of computational resources.

The optimization of illumination patterns is carried out using a data-driven methodology, where a small number of training images can lead to near-perfect reconstruction capabilities. Ex-

tensive simulation results demonstrate that our method outperforms existing approaches, providing substantial improvements in both accuracy and speed.

Impact and Future Directions. The data-driven enhancements in signal processing detailed here not only improve the technical capabilities of deep networks but also offer significant practical benefits. For instance, in medical imaging, enhanced phase retrieval can lead to better diagnostic capabilities, while in telecommunications, optimized signal recovery can increase the efficiency and reliability of data transmission.

Looking forward, the integration of advanced machine learning techniques with traditional signal processing tasks opens up exciting avenues for research and application. Continued advancements in this field are expected to drive further improvements in computational efficiency and solution accuracy, paving the way for new innovations in AI-enabled technologies.

Overall, these enhancements in signal processing exemplify how data-driven approaches can fundamentally transform the capabilities of deep networks, aligning with the broader goal of advancing the field of artificial intelligence towards more efficient, accurate, and applicable solutions across a wide range of disciplines.

1.4 Glimpse of approaches in this thesis

My research agenda centers on **exposing vulnerabilities in AI systems and meaningfully enhancing their trustworthiness**. By systematically uncovering flaws, I raise awareness of the precautions necessary for utilizing AI in high-stakes scenarios. My overarching mission is to eliminate dangerous deficiencies and develop robust intelligent machines that the public can confidently rely on. My methodology involves identifying vulnerabilities, quantifying worst-case perfor-

mance via attacks, and generalizing insights to practical deployment settings. I further investigate techniques to verifiably strengthen model resilience against common real-world manipulations. The goal is rigorous AI safety research that steers progress towards responsible and beneficial systems.

For highly effective black-box attacks, we developed ensemble-based and context-aware approaches. My method **optimizes over ensemble model weight spaces [64, 65]** to craft adversarial examples with extreme efficiency - over 30 times faster than existing input space attacks. Through multi-modal testing, I've shown that simultaneous attacks in this compact space can fool systems on diverse tasks using outputs only. As ensembles incorporate diverse models, their universality and attack potency increases, underscoring the need to evaluate deployment robustness against such attacks. This offers a generalizable methodology to surface vulnerabilities. Furthermore, I've also **weaponized context to mount impactful attacks [66, 63]**, manipulating the statistical relationships and invariants models implicitly rely on. By traversing contextual graphs encoding these dependencies, I introduce physically plausible inconsistencies that profoundly mislead systems. Despite divergence from training data, cross-dataset effectiveness arises from real-world underpinnings. My approach boosts black-box attack success over 20 percentage points with few queries, spotlighting reasoning vulnerabilities. As a defense mechanism, we **harness context as a detection mechanism [525]**. Noting language models' aptitude for encoding environmental plausibility, we developed techniques to perform model-agnostic consistency checks. By estimating scene likelihood under a language prior, incongruous object configurations crafted to deliberately mislead systems are identifiable despite model divergences. Our methodology highlights that progressing intelligence implies deeper reasoning interdependence; while contextual relationships empower inference, incoherencies within can profoundly undermine systems.

To protect user privacy, I developed an algorithm that can seamlessly **de-identify facial images while retaining utility for downstream tasks [60]**. Unlike previous approaches, mine is grounded in differential privacy and ensemble learning. I extract identity features then replace them with synthesized disguises using variational methods, maximizing obfuscation and non-invertibility to prevent re-identification. Additionally, I leverage supervision signals from an ensemble model to disentangle identity attributes from utility attributes like expressions.

To enhance the robustness and efficiency of computational imaging pipelines, including Fourier phase retrieval and coded diffraction imaging, I developed a framework that learns reference signals or illumination patterns using a small number of training images. This framework employs an unrolled network as a solver. Once learned, the reference signals or illumination patterns serve as priors, significantly improving the efficiency of signal reconstruction [62, 213, 217, 61].

1.5 Thesis organization

In this chapter, i.e. chapter 1, I introduce the motivation of my research, the problem of interests, how such problems are addressed in the literature, and a brief description of my solutions.

From chapter 2 to chapter 5, I will **revealed the vulnerabilities of neural networks [66, 64, 63, 65]** by demonstrating several efficient blackbox attacks on diverse deep learning models over different tasks. In chapter 6, I focus on **enhancing user privacy [60]** by innovating facial de-identification methods. In chapter 7 and 8, I present methods [62, 213, 217, 61] to enhance the robustness and efficiency of computational imaging systems.

Chapter 2

Blackbox Attacks via Surrogate Ensemble Search

2.1 Introduction

Deep neural network (DNN) models are known to be vulnerable to adversarial attacks [438, 165, 372, 373]. Many methods have been proposed in recent years to generate adversarial attacks [165, 261, 326, 131, 508, 300, 208, 319] (or to defend against such attacks [326, 455, 514, 178, 334, 308, 409, 507, 395, 25]). Attack methods for blackbox models can be divided into two broad categories: transfer- and query-based methods. Transfer-based methods generate attacks for some (whitebox) surrogate models via backpropagation and test if they fool the victim models [372, 373]. They are usually agnostic to victim models as they do not require or readily use any feedback; and they often provide lower success rates compared to query-based methods. On the other hand, query-based attacks achieve high success rate but at the expense of querying the victim model

several times to find perturbation directions that reduce the victim model loss [92, 459, 177, 278, 219]. One possible way to achieve a high success rate while keeping the number of queries small, is to combine the transfer and query attacks. While there has been impressive recent work along this direction [101, 208, 445, 319], the state-of-the-art methods [445, 319] still require hundreds of or more queries to be successful at targeted attacks. Such attacks are infeasible for limited-access settings where a user cannot query a model that many times[163].

Given this premise, we design a new method for blackbox attacks via surrogate ensemble search (BASES), combining transfer and query ideas, to fool a given victim model with higher success rates and fewer queries compared to state-of-the-art methods. For example, our evaluation shows that BASES (on average) only requires 3 queries per image to achieve over a 90% success rate for targeted attacks, which is at least $30\times$ fewer queries compared to state-of-the-art methods [208, 319]. BASES consists of two key steps that can be viewed as bilevel optimization steps. 1) A perturbation machine generates a query for the victim model based on weights assigned to the surrogate models. 2) The victim model's feedback is used to change weights of the perturbation machine to refine the query. Figure 7.1 depicts these steps.

We first define a perturbation machine (PM) that generates a single perturbation to fool all the (whitebox) models in the surrogate ensemble. We use a surrogate ensemble for two reasons: 1) It is known to provide better transfer attacks [309, 131]. The assumption is that if an adversarial image can fool multiple surrogate models, then it is very likely to fool a victim model as well. For the same reason, an ensemble with different and diverse surrogate models provides better attack transfer. 2) Our main interest is in searching for perturbations that can fool the given victim model. A single surrogate model provides a fixed perturbation; hence, it does not offer flexibility to search

over perturbations. To facilitate search over perturbations, we define the adversarial loss for the PM as a function of weights assigned to each model in the ensemble. By changing the weights of the loss function, we can generate different perturbations and steer in a direction that fools the victim model. It is worth noting that perturbations generated by a surrogate ensemble with an arbitrary set of weights often fools all the surrogate models, but they do not guarantee success on unseen victim models; therefore, searching over the weights space for surrogate models is necessary.

Since the number of models in the surrogate ensemble is small, the search space is low dimensional and requires extremely small number of queries compared to other query-based approaches. In our method, we further simplify the search process by updating one weight element at a time, which is equivalent to coordinate descent, which has been shown to be effective in query-based attacks [92, 177]. Since it searches in orthogonal directions instead of estimating the full gradients, it is query efficient. This strategy requires 2 queries per coordinate update but offers success rates as good as that given by performing a full gradient update step (as shown in Section 2.4). Reducing the dimension of the search space while maintaining high success rate for query-based attacks is an active area of research [177, 208, 445, 319], and our proposed method pushes the boundary in this area.

We perform extensive experiments for (score-based) blackbox attacks using a variety of surrogate and blackbox victim models for both targeted and untargeted attacks. We select PyTorch Torchvision [378] as our model zoo, which contains 56 image classification models trained on ImageNet [120] that span a wide range of architectures. We demonstrate superior performance by a large margin over state-of-the-art approaches, especially for targeted attacks. Furthermore, we tested

the perturbations generated by our method for attacks on hard-label classifiers. Our results show that the perturbations generated by our method are highly transferable.

The main contributions of this paper are as follows.

- We propose a novel, yet simple method, BASES, for effective and query-efficient blackbox attacks. The method adjusts weights of the surrogate ensemble by querying the victim model and achieves high fooling rate targeted attack with a very small number of queries.
- We perform extensive experiments to demonstrate that BASES outperforms state-of-the-art methods [101, 208, 445, 319] by a large margin; over 90% targeted success rate with less than 3 queries, which is at least $30\times$ fewer than other method.
- We also demonstrate the effectiveness under a real-world blackbox setting by attacking Google Cloud Vision API and achieve 91% untargeted fooling rate with 2.9 queries ($3\times$ less than [208]).
- The perturbations from BASES are highly transferable and can also be used for hard-label attacks. In this challenging setting, we can achieve over 90% fooling rate for targeted and almost perfect fooling rate for untargeted attacks on a variety of models using less than 3 and 2 queries, respectively.

2.2 Related work

Ensemble-based transfer attacks. Transferable adversarial examples that can fool one model can also fool a different model [372, 373, 309, 281] Transfer-based untargeted attacks are considered 'easy' since the adversarial examples can disrupt feature extractors into unrelated directions (e.g., in MIM [131], the fooling rate for some models can be as high as 87.9%). In contrast, transfer-based targeted attacks often suffer from low fooling rates (e.g., MIM shows a transfer rate of about 20%

at best). To improve the transfer rate, several methods use ensemble based approach. To combine the information from different surrogate models, [309] fuses probability scores, and [131] proposes combining logits. While these methods have been effective, the most natural and generic approach is to combine losses, which can be used for tasks beyond classification [86, 500]. MGAA [528] iteratively selects a set of surrogate models from an ensemble, to perform meta train and meta test steps to shrink the gap between whitebox and blackbox gradient directions. Previous ensemble approaches typically assign equal weights for each surrogate model. In contrast, we update weights for different surrogate models based on the victim model feedback.

Query-based attacks. Unlike transfer-based attacks, query-based attacks do not make assumptions that surrogate models share similarity with the victim model. They can usually achieve high fooling rates even for targeted attacks (but at the expense of queries) [92, 219, 459]. The query complexity is proportional to the dimension of the search space. Queries over the entire image space can be extremely expensive [92], requiring millions of queries for targeted attack [459]. To reduce the query complexity, a number of approaches have attempted to reduce the search space dimension or leverage transferable priors or surrogate models to generate queries. SimBA-DCT [177] searches over the low DCT frequencies. P-RGF [101] utilizes surrogate gradients as a transfer-based prior, and draws random vectors from a low-dimensional subspace for gradient estimation. TREMBA [208] trains a perturbation generator and traverses over the low-dimensional latent space. ODS [445] optimizes in the logit space to diversify perturbations for the output space. GFCS [319] searches along the direction of surrogate gradients, and falls back to ODS if surrogate gradients fail. We summarize the typical search space and average number of queries for some state-of-the-art methods in Table 2.1. In our approach, we further shrink the search dimension to as low as the number of

models in the ensemble. Since our search space is dense with adversarial perturbations, we show that a moderate-size ensemble with 20 models can generate successful targeted attacks for a variety of victim models while requiring only 3 queries (on average), which is at least 30 time fewer than that of existing methods.

2.3 Method

2.3.1 Preliminaries

We use additive perturbation [438, 165, 326] to generate a perturbed image as $x^\star = x + \delta$, where δ denotes the perturbation vector of same size as input image x. To ensure that the perturbation is imperceptible to humans, we usually constrain its ℓ_p norm to be less than a threshold, i.e., $\|\delta\|_p \leq \varepsilon$, where p is usually chosen from $\{2, \infty\}$. Such adversarial attacks for a victim model f can be generated by minimizing the so-called adversarial loss function $\mathcal{L}$ over δ such that the output $f(x + \delta)$ is as close to the desired (adversarial) output as possible. Specifically, the attack generator function maps the input image x to an adversarial image $x^\star$ such that the output $f(x^\star)$ is either far/different from the original output y for untargeted attacks, or close/identical to the desired output $y^\star$ for targeted attacks.

Let us consider a multi-class classifier $f(x) : x \mapsto z$, where $z = [z_1, \ldots, z_C]$ represents a logit vector at the last layer. The logit vector can be converted to a probability vector $p = \text{softmax}(z)$. We refer to such a classifier as a "score-based" or "soft-label" classifier. In contrast, a "hard-label" classifier provides a single label index out of a total of C classes. We can derive the hard label from the soft labels as $y = \arg\max_c f(x)_c$. For untargeted attacks, the objective is

to find $x^\star$ such that $\arg\max_c f(x^\star)_c \neq y$. For targeted attacks, the objective is to find $x^\star$ such that $\arg\max_c f(x^\star)_c = y^\star$, where $y^\star$ is the target label.

Many efforts on adversarial attacks use iterative variants of the fast signed gradient method (FGSM) [165] because of their simplicity and effectiveness. Notable examples include I-FGSM [261], PGD [326], and MIM [131]. We use PGD attack in our PM, which iteratively optimizes perturbations as

$$\delta^{t+1} = \Pi_\varepsilon \left(\delta^t - \lambda \, \mathbf{sign}(\nabla_\delta \mathcal{L}(x + \delta^t, y^\star)) \right), \tag{2.1}$$

where $\mathcal{L}$ is the loss function and Π_ε denotes a projection operator. There are many loss functions suitable for crafting adversarial attacks. We mainly employ the following margin loss, which has been shown to be effective in C&W attacks [79]:

$$\mathcal{L}(f(x), y^\star) = \max \left(\max_{j \neq y^\star} f(x)_j - f(x)_{y^\star}, -\kappa \right), \tag{2.2}$$

where κ is the margin parameter that adjusts the extent to which the example is 'adversarial.' A larger κ corresponds to a lower optimization loss. One advantage of C&W loss function is that its sign directly indicates whether the attack is successful or not ($+ve$ value indicates failure, $-ve$ value indicates success). Cross-entropy loss is also a popular loss function to consider, which has similar performance as margin loss (comparison results provided in supplementary material).

2.3.2 Perturbation machine with surrogate ensemble

Controlled query generation with PM. We define a perturbation machine (PM) to generate queries for the victim model as shown in Figure 7.1. The PM accepts an image and generates a perturbation to fool all the surrogate models. Furthermore, we seek some control over the perturbations generated by the PM to steer them in a direction that fools the victim model. To achieve these goals, we

construct the PM such that it minimizes a weighted adversarial loss function over the surrogate ensemble.

Adversarial loss functions for ensembles. Suppose our PM consists of N surrogate models given as $\mathcal{F} = \{f_1, \ldots, f_N\}$, each of which is assigned a weight in $\mathbf{w} = [w_1, \ldots, w_N]$ such that $\sum_{i=1}^{N} w_i = 1$. For any given image x and $\mathbf{w}$, we seek to find a perturbed image $x^\star(\mathbf{w})$ that fools the surrogate ensemble. Below we discuss three possible weighted ensemble loss functions-based optimization problems for targeted attack. Loss functions for untargeted attack can be derived similarly.

$$\textbf{weighted probabilities} \qquad x^\star(\mathbf{w}) = \arg\min_x -\log\left(\mathbf{1}_{y^\star} \cdot \sum_{i=1}^{N} w_i \,\text{softmax}(f_i(x))\right), \tag{2.3}$$

$$\textbf{weighted logits} \qquad x^\star(\mathbf{w}) = \arg\min_x \; \mathcal{L}\left(\sum_{i=1}^{N} w_i f_i(x), y^\star\right), \tag{2.4}$$

$$\textbf{weighted loss} \qquad x^\star(\mathbf{w}) = \arg\min_x \; \sum_{i=1}^{N} w_i \mathcal{L}(f_i(x), y^\star). \tag{2.5}$$

$y^\star$ denotes the target label and $\mathbf{1}_{y^\star}$ denotes its one-hot encoding. $\mathcal{L}$ represents some adversarial loss function (e.g., C&W loss). The first problem in (2.3) is the minimization of the softmax cross-entropy loss defined on the weighted combination of probability vectors from all ensemble models [309]. The second problem in (2.4) uses adversarial loss on the weighted combination of logits from the models [131], as the argument for the optimization. The third problem in (2.5) optimizes a weighted combination of adversarial losses over all models. The weighted loss formulation is the simplest and most generic ensemble approach that works not only for the classification task with logit or probability vectors, but also other tasks (e.g., object detection, segmentation) as long as the model losses can be aggregated [500]. Here, we focus on the weighted loss formulation, since it shows superior performance compared to weighted probabilities and logits formulations in our experiments (results in supplementary material).

Algorithm 1 presents a pseudocode for the PM module for a fixed set of weights. The PM accepts an image x and weights $\mathbf{w}$ along with the surrogate ensemble and returns the perturbed image $x^\star = x + \delta$ after a fixed number of signed gradient descent steps for the ensemble loss.

Algorithm 1 Perturbation Machine: $\delta, x^\star(\mathbf{w}) = \mathbf{PM}(x, \mathbf{w}, \delta_{\text{init}})$

Input:

Input x and the target class $y^\star$ (for untargeted attack $y^\star \neq y$);

Surrogate ensemble $\mathcal{F} = \{f_1, f_2, ..., f_N\}$;

Ensemble weights $\mathbf{w} = \{w_1, w_2, ..., w_N\}$;

Initial perturbation δ_{init}; Step size λ; Perturbation norm (ℓ_2/ℓ_∞) and bound ε

Output: Adversarial perturbation $\delta, x^\star(\mathbf{w})$

$\delta = \delta_{\text{init}}$

for $t = 1$ to T **do**

  Calculate $\mathcal{L}_{\mathbf{ens}} = \sum_{i=1}^{N} w_i \mathcal{L}_i(x + \delta, y^\star)$ ▷ *Ensemble loss*

  Update $\delta \leftarrow \delta - \lambda \cdot \mathbf{sign}(\nabla_\delta \mathcal{L}_{\mathbf{ens}})$ ▷ *Gradient of ensemble via backpropagation*

  Project $\delta \leftarrow \Pi_\varepsilon(\delta)$ ▷ *Project to the feasible set of ℓ_∞ or ℓ_2 ball*

end for

$x^\star(\mathbf{w}) \leftarrow x + \delta$

return $\delta, x^\star(\mathbf{w})$

2.3.3 Surrogate ensemble search as bilevel optimization

Let us assume that we are given a blackbox victim model, $f_{\mathbf{v}}$, that we seek to fool using a perturbed image generated by the PM (as illustrated in Figure 7.1). Suppose the adversarial loss for

the victim model is defined as $\mathcal{L}_{\mathbf{v}}$. To generate a perturbed image that fools the victim model, we want to solve the following optimization problem:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \ \mathcal{L}_{\mathbf{v}}(f_{\mathbf{v}}(x^{\star}(\mathbf{w})), y^{\star}). \tag{2.6}$$

The problem in (2.6) is bilevel optimization that seeks to update the weight vector $\mathbf{w}$ for the PM so that the generated $x^{\star}(\mathbf{w})$ fools the victim model. The PM in Algorithm 1 can be viewed as a function that solves the inner optimization problem in our bilevel optimization. The outer optimization problem searches over $\mathbf{w}$ to steer the PM towards a perturbation that fools the victim model.

BASES: blackbox attacks via surrogate ensemble search. Our objective is to maximize the attack success rate and minimize the number of queries on the victim model; hence, we adopt a simple yet effective iterative procedure to update the weights $\mathbf{w}$ and generate a sequence of queries. Pseudocode for our approach is shown in Algorithm 2. We initialize all entries in $\mathbf{w}$ to $1/N$ and generate the initial perturbed image $x^{\star}(\mathbf{w})$ for input x. We stop if the attack succeeds for the victim model; otherwise, we update $\mathbf{w}$ and generate a new set of perturbed images. We follow [92] and update $\mathbf{w}$ in a coordinate-wise manner, where at every outer iteration, we select nth index and generate two instances of $\mathbf{w}$ as $\mathbf{w}^{+}, \mathbf{w}^{-}$ by updating w_n as $w_n + \eta, w_n - \eta$, where η is a step size. We normalize the weight vectors so that the entries are non-negative and add up to 1. We generate perturbations $x^{\star}(\mathbf{w}^{+}), x^{\star}(\mathbf{w}^{-})$ using the PM and query the victim model. We compute the victim loss (or score) for $\{\mathbf{w}, \mathbf{w}^{+}, \mathbf{w}^{-}\}$ and select the weights, the perturbation vector, and the perturbed images corresponding to the smallest victim loss. We stop if the attack is successful with any query.

2.4 Experiments on classification

2.4.1 Experiment setup

Surrogate and victim models. We present experiments for blackbox attacks using image classification models from Pytorch Torchvision [378], which is a comprehensive and actively updated package for computer vision tasks. At the time of writing this paper, Torchvision offers 56 classification models trained on ImageNet dataset [120]. These models have different architectures and include the family of VGG [428], ResNet [187], SqueezeNet [218], DenseNet[202], ResNeXt [510], MobileNet [410, 197], EfficientNet [442], RegNet [387], Vision Transformer [134], and ConvNeXt [315]. We choose different models as the victim blackbox models for our experiments, as shown in Figures 2.2, 2.3, and 2.4. To construct an effective surrogate ensemble for the PM, we sample 20 models from different families: {`VGG-16-BN`, `ResNet-18`, `SqueezeNet-1.1`, `GoogleNet`, `MNASNet-1.0`, `DenseNet-161`, `EfficientNet-B0`, `RegNet-y-400`, `ResNeXt-101`, `Convnext-Small`, `ResNet-50`, `VGG-13`, `DenseNet-201`, `Inception-v3`, `ShuffleNet-1.0`, `MobileNet-v3-Small`, `Wide-ResNet-50`, `EfficientNet-B4`, `RegNet-x-400`, `VIT-B-16`}. We vary our ensemble size $N \in \{4, 10, 20\}$ by picking the first N model from the set. In most of the experiments, our method uses $N = 20$ models in the **PM**, unless otherwise specified. To validate the effectiveness of our methods in a practical blackbox setting, we also tested Google Cloud Vision API.

Comparison with other methods. We compare our method with some of the state-of-the-art methods for score-based blackbox attacks. TREMBA [208] is a powerful attack method that searches for perturbations by changing the latent code of a generator trained using a set of surrogate

models. GFCS [319] is a recently proposed surrogate-based attack method that probes the victim model using the surrogate gradient directions. We use their original code repositories [320, 209]. For completeness, we also compare with two earlier methods, ODS [445] and P-RGF [101], that leverage transferable priors, even though they have been shown to be less effective than GFCS and TREMBA. Additional details about comparison with TREMBA and GFCS are provided in the supplementary material.

Dataset. We evaluated all methods using 1000 ImageNet-like images from the NeurIPS-17 challenge [55, 262], which provides the ground truth label and a target label for each image.

Query budget. In this paper, we move towards a limited-access setting, since for many real-life applications, legitimate users will not be able to run many queries [163]. In contrast with TREMBA and GFCS, which set the maximum query count to $10,000$ and $50,000$, respectively, we set the maximum count to be 500 and only run our method for 50 queries in the worst case. (TREMBA also uses only 500 queries for Google Cloud Vision API to cut down the cost.)

Perturbation budget. We evaluated our method under both ℓ_∞ and ℓ_2 norm bound, with commonly used perturbation budgets of $\ell_\infty \leq 16$ and $\ell_2 \leq 255\sqrt{0.001D} = 3128$ on a 0–255 pixel intensity scale, where D denotes the number of pixels in the image. For attacking Google Cloud Vision API, we reduce the norm bound to $\ell_\infty \leq 12$ to align with the setting in TREMBA. Results for ℓ_2 norm bound are provided in the supplementary material.

Targeted vs untargeted attacks. All the methods achieve near perfect fooling rates for untargeted attacks in our experiments. This is because untargeted attack on image classifiers is not challenging [319], especially when the number of classes is large. Thus, we primarily report

experimental results on targeted attacks in the main text. Results for untargeted attacks are in the supplementary material.

2.4.2 Score-based attacks

Targeted attacks. Figure 2.2 presents a performance comparison of five methods for targeted attacks on three blackbox victim models. Our proposed method provides the highest fooling rate with the least number of queries. P-RGF is found to be ineffective (almost 0% success) for targeted attacks under low query budgets. TREMBA and GFCS are similar in performance; TREMBA shows better performance when query count is small, but GFCS matches TREMBA after nearly 100 queries. Nevertheless, our method clearly outperforms these two powerful methods by a large margin at any level of query count. We summarize the search space dimension $\mathcal{D}$ and query counts vs fooling rate of different methods under a limited (and realistic) query budget for both the targeted and untargeted attacks in Table 2.1. Our method is the most effective in terms of fooling rate vs number of queries (and has the smallest search dimension). *Additional results and details about fair comparison and fine tuning of TREMBA and GFCS are provided in the supplementary material.*

Surrogate ensemble size (N)**.** To evaluate the effect of surrogate ensemble size on the performance of our method, we performed targeted blackbox attacks experiment on three different victim models using three different sizes of surrogate ensemble: $N \in \{4, 10, 20\}$. The results are presented in Figure 2.3 in terms of fooling success rate vs number of queries. As we increase the ensemble size, the fooling rate also increases. With $N = 20$, the targeted attack fooling rate is almost perfect within 50 queries. Specifically, for `VGG-19` with $N = 20$, we improve from 54% success rate at the first query (with equal ensemble weights) to 96% success rate at the end of 50

Table 2.1: Number of queries vs fooling rate of different methods and the search space dimension $\mathcal{D}$.

Method	$\mathcal{D}$	Number of queries ($\mathbf{mean} \pm \mathbf{std}$) per image and fooling rate					
		VGG-19		DenseNet-121		ResNext-50	
		Targeted	Untargeted	Targeted	Untargeted	Targeted	Untargeted
P-RGF [101]	7,500	-	156 ± 113 93.5%	-	164 ± 112 92.9%	-	166 ± 116 92.5%
TREMBA [208]	1,568	92 ± 107 89.2%	2.4 ± 14 99.7%	70 ± 104 90.5%	5.9 ± 28 99.5%	100 ± 109 85.1%	7.5 ± 38 98.9%
ODS [445]	1,000	261 ± 125 49.0%	38 ± 48 99.9%	266 ± 123 49.7%	52 ± 64 99.0%	270 ± 116 42.7%	54 ± 65 98.4%
GFCS [319]	1,000	101 ± 95 89.1%	14 ± 21 100.0%	76 ± 75 95.2%	16 ± 36 99.9%	86 ± 87 92.9%	15 ± 18 99.7%
Ours	**20**	$\mathbf{3.0 \pm 5.4}$ **95.9%**	$\mathbf{1.2 \pm 2.4}$ **99.8%**	$\mathbf{1.8 \pm 2.7}$ **99.4%**	$\mathbf{1.2 \pm 1.8}$ **99.9%**	$\mathbf{1.8 \pm 2.6}$ **99.7%**	$\mathbf{1.2 \pm 0.9}$ **100.0%**

queries; this equals 78% improvement. `DenseNet-121` and `ResNext-50` can achieve 100% fooling rate with $N = 20$. With `DenseNet-121`, using 10 surrogate models, we can achieve a fooling rate of 98%. While using 4 models is challenging with respect to all victim models, we can see a rapid and significant improvement in fooling rates when the number of queries increases.

Comparison of whitebox (gradient) vs blackbox (queries). To check the effectiveness of our coordinate descent approach for updating $\mathbf{w}$, we compare its performance with the alternative approach of calculating the exact gradient of victim loss under the whitebox setting. The results are

presented in Figure 2.3 as dotted lines. We observe that our blackbox query approach provides similar results as the whitebox version, which implies the coordinate-wise update of $\mathbf{w}$ is as good as a complete gradient update.

2.4.3 Hard-label attacks

The queries generated by our PM are highly transferable and can be used to craft successful attacks for hard-label classifiers. To generate a sequence of queries for hard-label classifiers, we pick a 'surrogate victim' model and generate queries by updating $\mathbf{w}$ in the same manner as the score-based attacks for Q iterations (without termination). We store the queries generated at every iterations in a query set $\{\delta^1, \ldots, \delta^Q\}$. We test the victim hard-label blackbox model using $x + \delta$ by selecting δ from the set in a sequential order until either the attack succeeds or the queries finish.

In our experiments, we observed that this approach can achieve a high targeted attack fooling rate on a variety of models. We present the results of our experiment in Figure 2.4, where we report attack success rate vs query count for 6 models: {`MobileNet-V2`, `ResNet-34`, `ConvNeXt-Base`, `EfficientNet-B2`, `RegNet-x-8`, `VIT-L-16`}. We used `VGG-19` as the 'surrogate victim' model to generate the queries using the PM with 20 surrogate models. Using the saved surrogate perturbations, we can fool all models almost 100%, except for `VIT-L-16` [134] that is a vision transformer and architecturally very different from the majority of surrogate ensemble models (thus difficult to attack). Nevertheless, the fooling rate increases from $18\% \rightarrow 63\%$, which is a 250% improvement.

2.4.4 Attack on commercial Google Cloud Vision API

We demonstrate the effectiveness of our approach under a practical blackbox setting by attacking the Google Cloud Vision (GCV) label detection API. GCV detects and extracts information about entities in an image, across a very broad group of categories containing general objects, locations, activities, animal species, and products. Thus, the label set is very different from that of ImageNet, and largely unknown to us. We have no knowledge about the detection models in this API either. We randomly select 100 images from the aforementioned ImageNet dataset that are correctly classified by GCV, and perform untargeted attacks against GCV using 20 surrogate models with perturbation budget of $\ell_\infty \leq 12$ to align with the setting in TREMBA [208].

For each input image, GCV returns a list of labels, which are usually the top 10 labels ranked by probability. Under the success metric of changing the top 1 label to any other label, same as in [208], our attack can achieve a fooling rate of 91% with only 2.9 queries per image on average, which is much lower than 8 queries TREMBA reported for similar experiment. We present some successful examples in Figure 2.5. We present additional results in the supplementary material that show our attacks from classification can transfer to object detection models.

Comparison with TREMBA. TREMBA [208] requires one trained generator for each target class; thus, it is not feasible to test it for any arbitrary target label selected from 1000 classes in ImageNet. For a fair comparison, we attack each image using one of the 6 target labels available in trained TREMBA model $\{0, 20, 40, 60, 80, 100\}$ and average the query counts. Furthermore, TREMBA generator was trained using an ensemble of 4 surrogate models; while it is possible to train the generator with more surrogate models, training one generate per target label is expensive and non-trivial in terms of hyper-parameter tuning. Therefore, in our experiments, we used the

trained generator from the paper. It is worth pointing out that our method with 4 surrogate models (as shown in Figure 2.3) is still better than TREMBA in the low query count regime. TREMBA can provide better success rate at the expense of increased queries.

Why is our method better than TREMBA? TREMBA generates patterns by optimizing over the latent code of a trained generator, which contributes to the high success rate. TREMBA generator has a large enough range that it can generate adversarial perturbations that fool a victim model. Our experimental results suggests that the space of perturbations generated by our PM (via weighted surrogate ensemble) is better (in terms of diversity and low dimensionality) than TREMBA's generator. That is the reason why we see a steep slope for the first few queries in our success vs query curve.

Comparison with GFCS. To perform our experiments, we used the same set of $N = 20$ surrogate models for GFCS [319] that are used in our PM. GFCS used ℓ_2 norm constraint and did not compare with TREMBA. While our method can generate perturbations with ℓ_2 and ℓ_∞ constraints, TREMBA generates perturbations with ℓ_∞ constraint. To perform a fair comparison, we modified GFCS code to have ℓ_∞ constraint and tuned the hyper-parameters to achieve the best performance. The step-size is the key parameter that we choose as 0.005 after searching over a grid of $\{0.2, 0.02, 0.01, 0.005, 0.001, 0.0005\}$. As shown in 2.6, the performance reported in Figure 2.2 for ℓ_∞ attacks is on par with the performance achieved with original settings of ℓ_2 norm constraint.

Why is our method better than GFCS? Our method is more query efficient because we leverage all surrogate models for each query, whereas GFCS only uses one surrogate model per query. We can see that our method has the steepest slope in Figure 2.6 and the highest success at the starting point.

Note about P-RGF. The original implementation of P-RGF is in Tensorflow, but to unify the platform, we use the Pytorch implementation provided by GFCS [320].

Comparison with Simulator Attack. We use the same setting as in simulator attack [325] that tests 3 victim blackbox models {`DenseNet-121`, `ResNeXt-101 (32×4d)`, `ResNeXt-101 (64×4d)`} and uses 16 surrogate models {`VGG-11/13/16/19`, `VGG-11/13/16/19 (BN)`, `ResNet-18/34/50/101/152`, `DenseNet-161/169/201`}. All of these models are trained on TinyImageNet [405] dataset and we obtain the pretrained weights from [325]. We randomly select 1000 tinyImageNet images and use incremental target label selection for targeted attacks. Target label $y_{adv} = (y + 1) \mod C$, where y is the original label and total number of classes is $C = 200$. Perturbation budget for targeted attack is $\ell_2 \leq 4.6 \times 255 = 1173$, and for untargeted attack $\ell_\infty \leq 8$. As shown in Table 2.2, we achieve perfect fooling rates with less than two queries on average for both targeted and untargeted attacks. Specifically, for `ResNeXt-101 (32×4d)`, we achieve 100% targeted fooling rate with an average query count of 2.0, (min = 1, max = 26, median = 1). In contrast, simulator attack achieves 84.9% fooling rate using 2558 queries, which is $1279\times$ more expensive than ours. For untargeted attack, the trend is similar that our method is 811–$1445\times$ more query efficient than simulator attack.

Comparison with combining transfer and query-based attacks. Hybrid attack in [436] is one of the earliest works that combines transfer and query-based attacks. It uses surrogate models to generate the initial query, which is later updated using feedback from the blackbox victim model via pure query-based methods. To verify that our proposed method is advantageous, we use the perturbations generated by our ensemble models with equal weights as the initial query, and for every failed query we deploy a powerful pure query-based method square attack [13]. Following the

Table 2.2: Number of queries vs fooling rate of different methods on TinyImageNet dataset.

Method	Number of queries ($\mathbf{mean}/\mathbf{median}$) per image and fooling rate					
	DenseNet-121		ResNeXt-101 (32×4d)		ResNeXt-101 (64×4d)	
	Targeted	Untargeted	Targeted	Untargeted	Targeted	Untargeted
NES [219]	4625 / 4337	1306 / 510	4959 / 4703	2104 / 765	4758 / 4440	2078 / 816
	88.5%	74.3%	88.0%	45.3%	88.2%	45.5%
Meta [135]	5420 / 5506	3789 / 3202	5440 / 5249	4101 / 3712	5661 / 5250	4012 / 3649
	24.2%	71.1%	21.0%	33.8%	18.2%	36.0%
Bandits [220]	2724 / 1860	964 / 520	3550 / 2700	1737 / 954	3542 / 2854	1662 / 1014
	85.1%	99.2%	72.2%	94.1%	72.4%	95.3%
Simulator [325]	1959 / 1399	811 / 431	2558 / 1966	1380 / 850	2488 / 1982	1445 / 878
	89.8%	99.4%	84.9%	96.8%	83.9%	97.9%
Ours	**1.5 / 1**	**1.0 / 1**	**2.0 / 1**	**1.0 / 1**	**2.0 / 1**	**1.0 / 1**
	100.0%	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

same setting as in our Table 2.1 and Figure 2.2, we perform targeted attack on `DenseNet-121` with a perturbation budget of $\ell_\infty \leq 16$. The transfer rate of initial perturbed images is 75.5%. We attack the remaining 24.5% failed perturbed images using square attack by allowing a maximum query count of 500 (same setting as other baseline methods). On this subset of images, we observed a 33.9% fooling rate with query count ($\mathbf{mean} \pm \mathbf{std}$): 238.2 ± 127.4. Overall, including the images that can initially transfer, the combination of [436] and [13] achieves a fooling rate of 83.8%, with query count ($\mathbf{mean} \pm \mathbf{std}$): 24.5 ± 81.4. In comparison, our method achieves a 99.4% fooling rate with a query count of 1.8 ± 2.7. Similar trends appear for other victim models, as shown in Table

Table 2.3: Number of queries vs fooling rate for hybrid methods that combine transfer and query-based attacks.

Models	Fooling rate and number of queries ($\mathbf{mean} \pm \mathbf{std}$) per image	
	Combine [436] and [13]	**Ours**
VGG-19	64.7% ; 34.1 $\pm$ 99.2	**95.9% ; 3.0 $\pm$ 5.4**
DenseNet-121	83.8% ; 24.5 $\pm$ 81.4	**99.4% ; 1.8 $\pm$ 2.7**
ResNext-50	84.3% ; 24.7 $\pm$ 81.4	**99.7% ; 1.8 $\pm$ 2.6**

2.3. Our main takeaway is that even though the surrogate ensemble provides highly transferable perturbation or perturbations that can be used as initialization for query-based optimization methods. The query-based methods lose their advantage by querying over a high dimensional image space. Our method searches over the weights of the ensemble loss, which is very low dimension and provides query efficiency.

Untargeted attacks. Un-targeted attacks are 'easy' [319] in image classification, especially when the number of classes is large (e.g., in ImageNet that has 1000 categories). We show that our method can readily achieve a fooling rate over 99% with only 1–2 queries (on average), as depicted in Figure 2.7 below and Table 2.1 in the main text. The initial perturbations from the PM (with all ensemble weights set to $1/N$) can already achieve a fooling rate of over 94%, close to that of TREMBA. Other methods require tens or hundreds of queries to achieve near-perfect success rate.

2.4.5 Classification performance on clean images

To ensure that all the models provide reasonably correct classification results on clean images, we calculate the classification accuracy of all ImageNet models on the 1000 test images. Our calculation shows that they have a Top-1 accuracy of ($\mathbf{mean} \pm \mathbf{std}$): $89.1\% \pm 6.5\%$. Among all the models tested, `Convnext-Smal` achieves the highest accuracy at 96.8%, and `SqueezeNet-1.1` gets the lowest at 68.8%.

2.5 Experiments on object detection

To demonstrate the generalizability of BASES beyond classification tasks, we also performed experiments for vanishing attacks on object detectors. The results indicate that our proposed method can be easily adopted for other tasks.

2.5.1 Experiment setup

Surrogate and victim models. We evaluate BASES using object detectors from MMDetection [90, 89], which provides a diverse set of models form over fifty model families, including `Faster R-CNN [396]`, `YOLOv3 [393]`, `RetinaNet [301]`, `FreeAnchor [539]`, `RepPoints [521]`, `CenterNet [548]`, `DETR [76]`, and `Deformable DETR [552]`. We choose different models {`RetinaNet`, `RepPoints`, `Deformable DETR`} as victim blackbox models, as shown in Figure 2.9. For surrogate models in the PM, we select some popular models {`Faster R-CNN`, `YOLOv3`, `FreeAnchor`, `DETR`, `CenterNet`} and vary our ensemble size $N \in \{2, 3, 4, 5\}$ by choosing the first N models from the set.

Dataset, attacks, query, and perturbation budgets. All models are trained on COCO 2017 train dataset [302]. We randomly sample 100 images of stop sign from COCO 2014 validation dataset to perform blackbox vanishing attacks. The attack is considered successful if the victim model fails to detect the stop sign in the adversarial image. The constraints on the query budget $Q \leq 50$ and perturbation budget $\ell_\infty \leq 16$ are the same as the classification setting.

Loss functions and ensemble loss. For individual surrogate models, we use the original loss function used for their training. We defined the ensemble loss as a weighted combination of loss over all the surrogate models. The confidence score of stop sign detected by the victim model is used as a feedback from the victim model.

2.5.2 Attacks on object detection

The results of attacking object detectors are shown in Figure 2.9 and Table 2.4. We observe that our attack method is effective and query efficient in attacking object detectors. In particular, for RetinaNet, a simple transfer attack (first iteration) has 27% fooling rate with $N = 2$ surrogate models. The fooling rate improve from $27\% \rightarrow 81\%$ with a small number of queries, which is a 300% improvement. Our attack gets stronger as the number of surrogate models increases. When $N = 5$, we can get almost perfect ($\geq 99\%$) fooling rate for all victim models with less than 3 queries on average.

Table 2.4: Number of queries per image and fooling rate of attacks on three victim models using different number N of surrogate models in PM.

N	Fooling rate and number of queries ($\mathbf{mean} \pm \mathbf{std}$) per image		
	RetinaNet	RepPoints	Deformable DETR
2	81% ; 8.5 $\pm$ 11	86% ; 8.0 $\pm$ 9.9	74% ; 8.5 $\pm$ 11
3	100% ; 3.9 $\pm$ 6.5	99% ; 2.8 $\pm$ 4.1	95% ; 5.4 $\pm$ 9.3
4	100% ; 2.2 $\pm$ 2.4	98% ; 2.2 $\pm$ 3.1	97% ; 2.1 $\pm$ 2.2
5	100% ; 2.0 $\pm$ 2.1	99% ; 2.1 $\pm$ 3.0	99% ; 2.1 $\pm$ 2.9

2.5.3 Attacks on Google Cloud Vision API

We also observe that the attacks generated by our method can also fool object detection models, as shown in Figure 2.10.

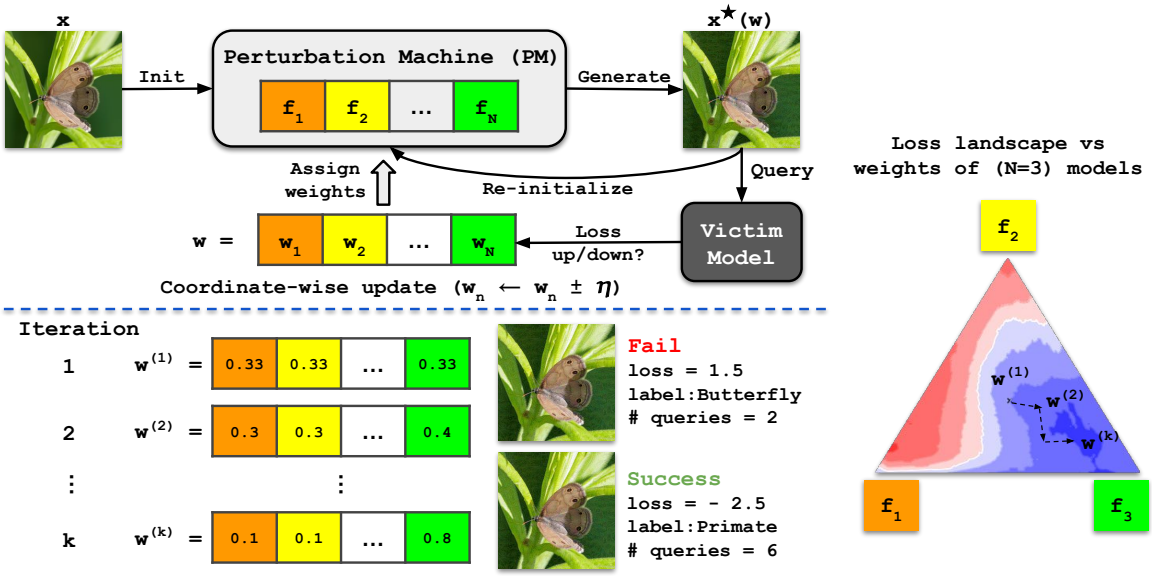


Figure 2.1: BASES for score-based attack. (**Top-left**) We define a perturbation machine (PM) using a fixed set of N surrogate models, each of which is assigned a weight value as $\mathbf{w} = [w_1, \ldots, w_N]$. The PM generates a perturbed image $x^\star(\mathbf{w})$ for a given input image x by minimizing the perturbation loss that is defined as a function of $\mathbf{w}$. To fool a victim model, we update one coordinate in $\mathbf{w}$ at a time while querying the victim model using $x^\star(\mathbf{w})$ generated by the PM. We can view this approach as a bi-level optimization or search procedure; the PM generates a perturbed image with the given weights $x^\star(\mathbf{w})$ in the inner level, while we update $\mathbf{w}$ in the outer level. (**Bottom-left**) We visualize weights and perturbed images for a few iterations. We stop as soon as the attack is successful (e.g. original label - 'Butterfly' is changed to target label - 'Primate' for targeted attack). (**Right**) Victim loss values for different weights along the Barycentric coordinates on the triangle. We start with equal weights (at the centroid) and traverse the space of $\mathbf{w}$ to reduce loss (concentrate on model f_3). Red color indicates large loss values (unsuccessful attack), and blue indicates low loss (successful attack).

Algorithm 2 BASES: Blackbox Attack via Surrogate Ensemble Search

Input:

Input x and the target class $y^\star$ (for untargeted attack $y^\star \neq y$); Victim model $f_{\mathbf{v}}$; Maximum number of queries Q; Learning rate η; Perturbation machine with surrogate ensemble

Output: Adversarial perturbation $\delta, x^\star$

Initialize $\delta = 0$; $q = 0$; $\mathbf{w} = \{\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}\}$

Generate perturbation via PM: $\delta, x^\star(\mathbf{w}) = \mathbf{PM}(x, \mathbf{w}, \delta)$ ▷ *first query with equal weights*

Query victim model: $z = f_{\mathbf{v}}(x + \delta)$

Update query count: $q \leftarrow q + 1$

if $\arg\max_c z_c = y^\star$ **then**

  break ▷ *stop if attack is successful*

end if

while $q < Q$ **do**

  Update surrogate ensemble weights as follows. ▷ *outer level updates weights*

  Pick a surrogate index n ▷ *cyclic or random order*

  Compute $\mathbf{w}^+, \mathbf{w}^-$ by updating w_n as $w_n + \eta, w_n - \eta$, respectively

  Generate perturbation $x^\star(\mathbf{w}^+), x^\star(\mathbf{w}^-)$ via PM ▷ *inner level generates query*

  Query victim model: $f_{\mathbf{v}}(x^\star(\mathbf{w}^+)), f_{\mathbf{v}}(x^\star(\mathbf{w}^-))$ ▷ *2 queries per coordinate*

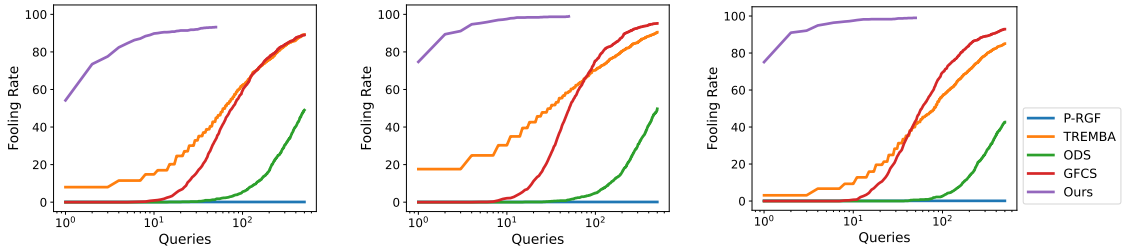
  Calculate victim model loss for $\{\mathbf{w}, \mathbf{w}^+, \mathbf{w}^-\}$ as $\mathcal{L}_{\mathbf{v}}(\mathbf{w}), \mathcal{L}_{\mathbf{v}}(\mathbf{w}^+), \mathcal{L}_{\mathbf{v}}(\mathbf{w}^-)$

  Select $\mathbf{w}, \delta, x^\star(\mathbf{w})$ for the weight vector with the smallest loss

  Increment q after every query, and stop if the attack is successful for any query

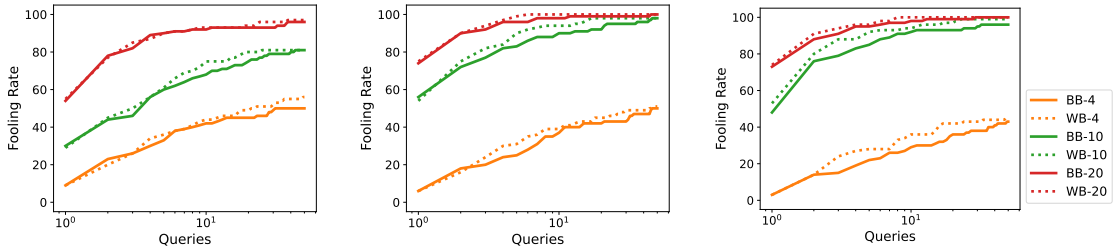
end while

return δ



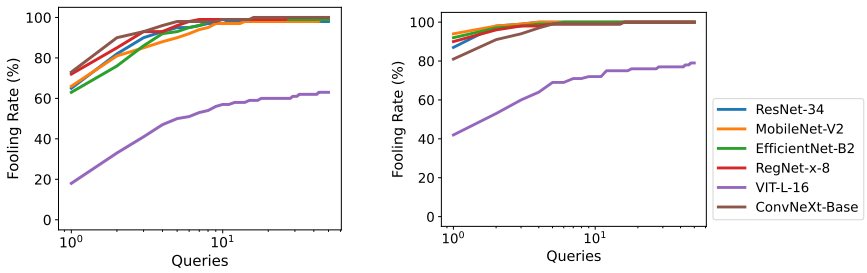
(a) VGG-19 **(b)** DenseNet-121 **(c)** ResNext-50

Figure 2.2: Comparison of 5 attack methods on three victim models under perturbation budget $l_\infty \leq 16$ for targeted attack. Our method achieves high success rate (over 90%) with few queries (average of 3 per image).



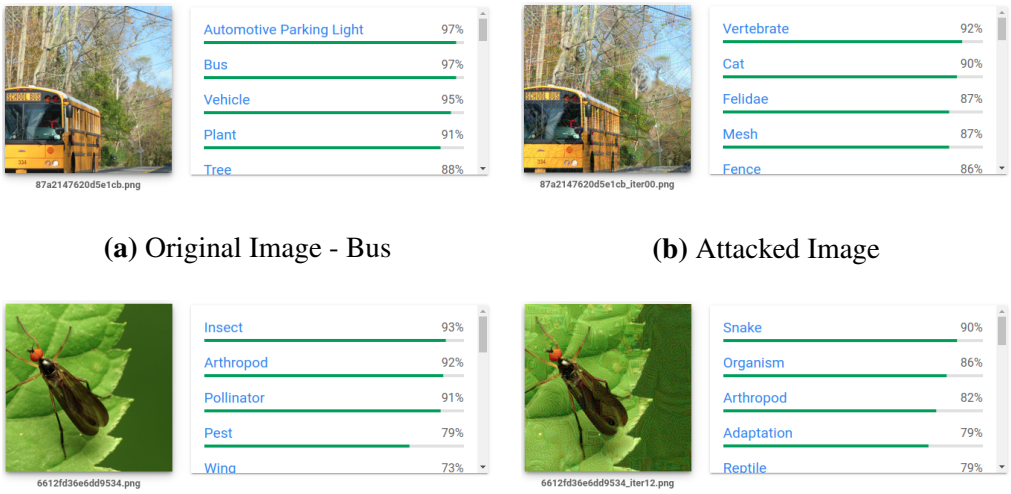
(a) VGG-19 **(b)** DenseNet-121 **(c)** ResNext-50

Figure 2.3: Comparison of targeted attack fooling rate with different number of ensemble models $N \in \{4, 10, 20\}$ in PM. Every experiment is performed with whitebox gradient (denoted as 'WB' with dotted lines) and blackbox score-based coordinate descent (denoted as 'BB' with solid lines). Experiment was run on 100 images.



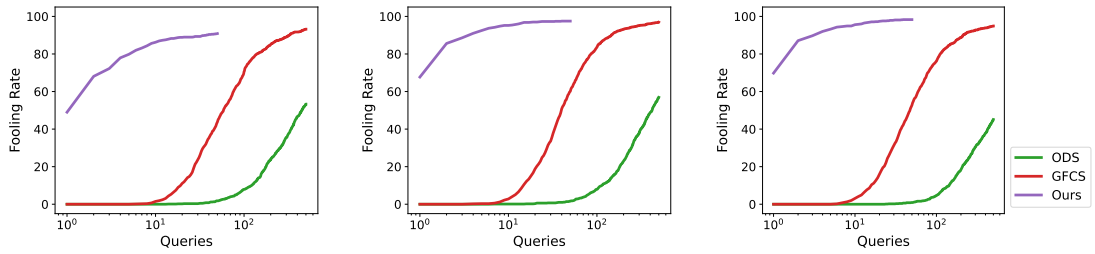
(a) Targeted **(b)** Untargeted

Figure 2.4: Performance of blackbox attack on 6 hard-label classifiers. Our method generates a sequence of queries for targeted attack using `VGG-19` as a victim model while the PM has $N = 20$ models in the surrogate ensemble. Experiment performed on 100 images.



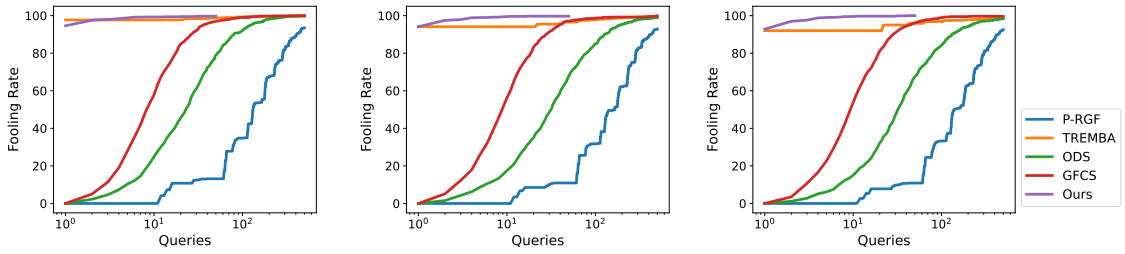
(c) Original Image - Fly **(d)** Attacked Image

Figure 2.5: Visualization of some successful attacks on Google Cloud Vision.



(a) VGG-19 **(b)** DenseNet-121 **(c)** ResNext-50

Figure 2.6: Adversarial attacks generated with ℓ_2 constraint (equivalent to Figure 2.2 in main text that uses ℓ_∞ constraints). Comparison of our method with GFCS / ODS on three victim models under perturbation budget $\ell_2 \leq 3128$ for targeted attacks.



(a) VGG-19 **(b)** DenseNet-121 **(c)** ResNext-50

Figure 2.7: Untargeted attacks (version of Figure 2.2 in the main text). Comparison of 5 attack methods on three victim models under perturbation budget $l_\infty \leq 16$ for untargeted attack. All methods can achieve near perfect success rate within 500 queries.

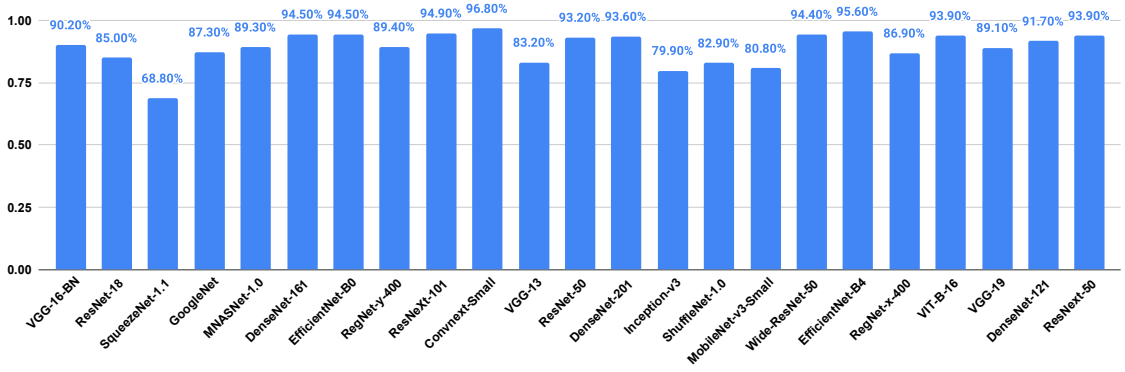
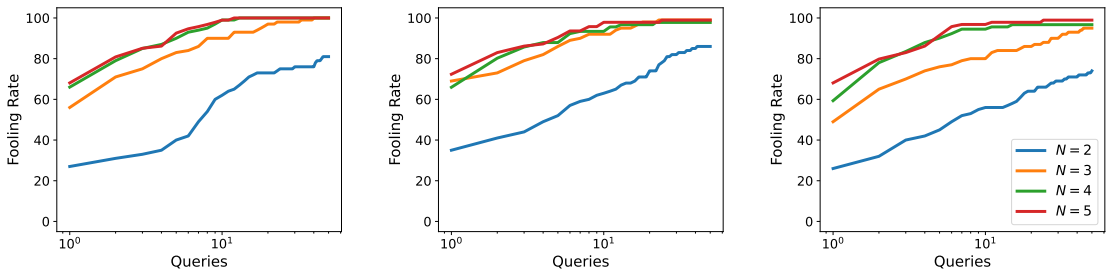


Figure 2.8: Top 1 classification accuracies of different ImageNet models used in our experiments.



(a) RetinaNet **(b)** RepPoints **(c)** Deformable DETR

Figure 2.9: Fooling rates for vanishing attacks on three victim object detectors using different number ($N \in \{2, 3, 4, 5\}$) of surrogate models in PM.

(a) Original Image - Bus **(b)** Attacked Image

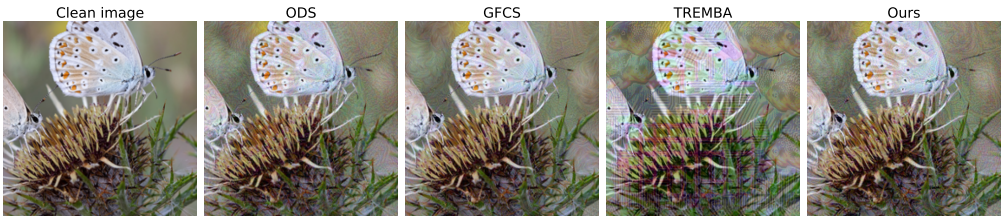
(c) Original Image - Fly **(d)** Attacked Image

Figure 2.10: Attacks generated by our PM can fool object detection models. Visualization of some successful attacks on Google Cloud Vision object detection API. (Compare to Figure 2.5 in main text.)

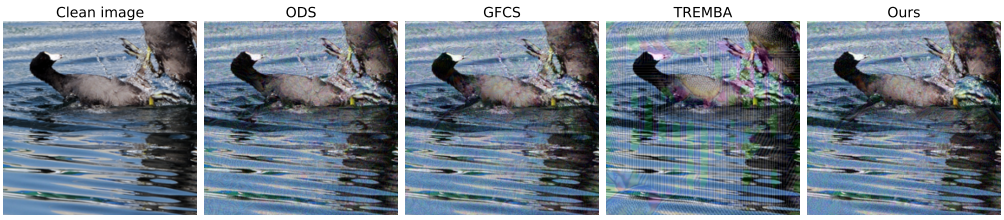
2.6 Visualization of adversarial examples

Classifiers. We present some examples of adversarial images generated by different methods for targeted attack on VGG-19 classifier in Figure 2.11. We observe that even with the same perturbation budget, $\ell_\infty \leq 16$, perturbation from our method is less visible than TREMBA, and is comparable with the ones from ODS and GFCS. TREMBA perturbs all images to 'Tench' and has a very structured semantic pattern that becomes visible. ODS, GFCS, and our method perturb 'Butterfly' to 'Dog', 'Coot' to 'Jacamar', and 'Parrot' to 'Fountain'.

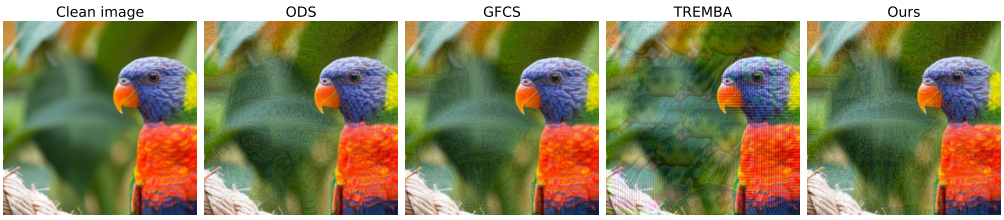
Detectors. We visualize some example images of attacking different object detectors in Figure 2.12. Our method effectively vanishes stop sign in the scene.



(a) Original: Butterfly; Target: Dog for ODS, GFCS, and Ours; Target: 'Tench' for TREMBA.



(b) Original: Coot; Target: Jacamar for ODS, GFCS, and Ours; Target: 'Tench' for TREMBA.



(c) Original: Parrot; Target: Fountain for ODS, GFCS, and Ours; Target: 'Tench' for TREMBA.

Figure 2.11: Visualization of adversarial images generated by different methods for targeted attack. (Corresponds to experiments in Figure 2.2 in main text.)

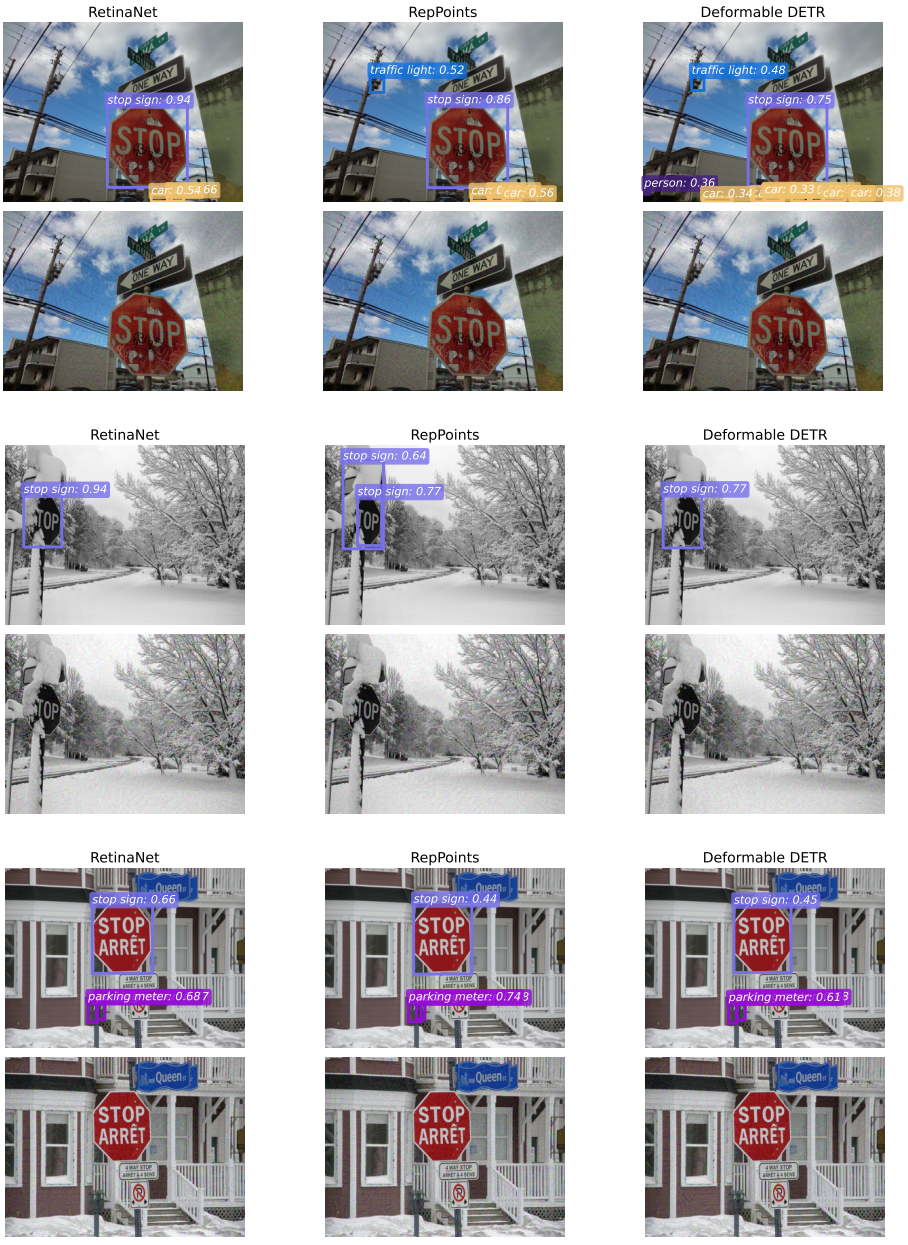


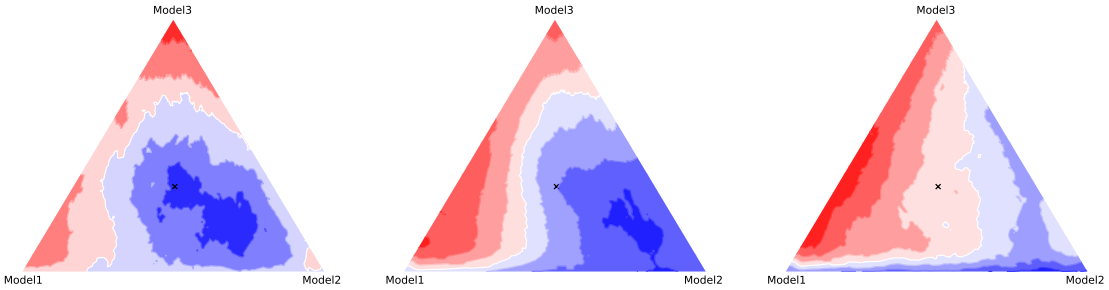
Figure 2.12: Visualization of adversarial images generated by different methods for vanishing attacks on 'stop sign'. Top row is detection on clean images and bottom row is detection on adversarial images. (Corresponds to results in Figure 2.9 with $N = 5$.)

2.6.1 Loss landscape vs ensemble weights

Why does ensemble weights-based query update work? We visualize the loss landscape of some victim models with respect to ensemble weights of three surrogate models in the PM. The plots in Figure 2.13 illustrate the loss, where the vertices of each triangle represent the surrogates models in the PM used for attacking a victim model on one image (as shown in sub-caption). The location of each point inside the triangle corresponds to the weight vector $\mathbf{w}$ (in terms of Barycentric coordinates). For instance, the centroid (marked by $\times$) has the barycentric coordinate $\mathbf{w} = [13, 13, 13]$, which implies the losses for all the surrogate models in the ensemble are weighted equally. More weight is given to a model if the weight vector moves closer to the vertex of that model. The color of each point inside the triangle represents the victim loss value for the corresponding $\mathbf{w}$. The attack is more successful when the loss value is low (indicated by blue color) and less successful when the loss value is high (indicated by red color). We created this figure using VGG-16, ResNet-18, and SqueezeNet as Model 1,2, and 3, respectively. The main takeaway is that, in many cases, an arbitrary weight vector does not provide successful perturbation for a given victim model; therefore, we need to adjust the weights to generate successful attacks.

2.7 Conclusion and discussion

We propose a novel and simple approach, BASES, to effectively perform blackbox attacks in a query-efficient manner, by searching over the weight space of ensemble models. Our extensive experiments demonstrate that a wide range of models are vulnerable to our attacks at the fooling rate of over 90% with as few as 3 queries for targeted attacks. The attacks generated by our method are



(a) VGG-19, Image A **(b)** ResNet-34, Image A **(c)** ResNet-34, Image B

Figure 2.13: Illustration of the effect of weights of ensemble models on the attack loss for a victim model. Red color indicates large loss values (unsuccessful attack), and blue indicates small loss (successful attack).

highly transferable and can also be used to attack hard-label classifiers. Attacks on Google Cloud Vision API further demonstrates that our attacks are generalizable beyond the surrogate and victim models in our experiments.

Limitations. 1) Our method needs a diverse ensemble for attacks to be successful. Even though the search space is low-dimensional, the generated queries should span a large space so that they can fool any given victim model. This is not a major limitation for image classification task as a large number of models are available, but it can be a limitation for other tasks. 2) Our method relies on the PM to generate a perturbation query for every given set of weights. The perturbation generation over surrogate ensemble is computationally expensive, especially as the ensemble size becomes large. In our experiments, one query generation with $\{4, 10, 20\}$ surrogate models requires nearly $\{2.4s, 9.6s, 18s\}$ per image on Nvidia GeForce RTX 2080 TI. Since our method requires a small number of queries, the overall computation time of our method remains small.

Societal impacts. We propose an effective and query efficient approach for blackbox attacks. Such adversarial attacks can potentially be used for malicious purposes. Our work can help further explain the vulnerabilities of DNN models and reduce technological surprise. We also hope this work will motivate the community to develop more robust and reliable models, since DNNs are widely used in real-life or even safety-critical applications.

Chapter 3

Ensemble-based Blackbox Attacks on Dense Prediction

3.1 Introduction

Computer vision models (e.g., classification, object detection, segmentation, and depth estimation) are known to be vulnerable to carefully crafted adversarial examples [438, 165, 66, 172, 102]. Creating such adversarial attacks is easy for whitebox models, where the victim model is completely known [165, 261, 326, 131, 508]. In contrast, creating adversarial attacks for blackbox models, where the victim model is unknown, remains a challenging task [309, 506, 16]. Most of the existing blackbox attack methods have been developed for classification models [319, 208, 101, 445]. Blackbox attacks for dense prediction models such as object detection and segmentation are relatively less studied [66, 172, 297], and most of the existing ones mainly focus on untargeted attacks [172]. Furthermore, a vast majority of these methods are based on transfer attacks, in which

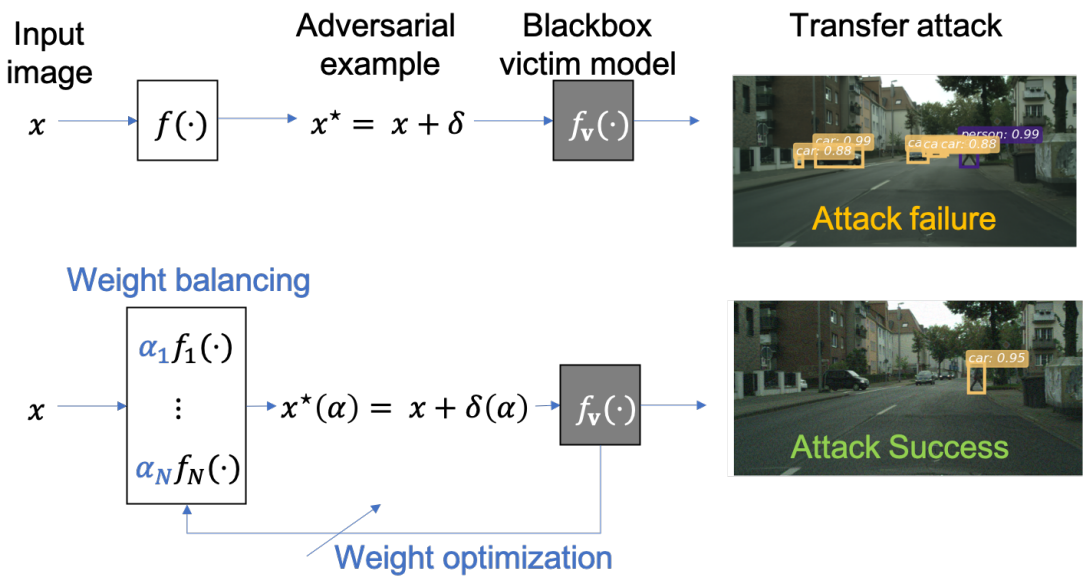


Figure 3.1: Illustration of the targeted ensemble-based blackbox attack. (Top) Attack generated by a single surrogate model does not transfer on the victim blackbox model (person does not map to car). (Bottom) Attack generated by weight balancing and optimization can transfer on a variety of victim models (person is mapped to car).

a surrogate (whitebox) model is used to generate the adversarial example that is tested on the victim model. However, the success rate of such transfer-based attacks is often low, especially for targeted attacks [208, 101, 445].

In this paper, we propose and evaluate an ensemble-based blackbox attack method for objection detection and segmentation. Our method is inspired by three key observations: 1) targeted attacks generated by a single surrogate model are rarely successful; 2) attacks generated by an ensemble of surrogate models are highly successful if the contribution from all the models is properly normalized; and 3) attacks generated by an ensemble for a specific victim model can be further improved by adjusting the contributions of different surrogate models. The overall idea of

the proposed work is illustrated in fig:intro. Our proposed method can be viewed as a combination of transfer- and query-based attacks, where we can adjust the contribution based on the feedback from the victim model using a small number of queries (5–20 in our experiments). In contrast, conventional query-based attacks require hundreds or thousands of queries from the victim model [92, 219, 459, 177].

We conduct comprehensive experiments to validate our proposed method and achieve state-of-the-art performance for both targeted and untargeted blackbox attacks on object detection. Specifically, our proposed method attains 29–53% success rate using only 5 queries for targeted attacks on object detectors, whereas the current state-of-the-art method [66] achieves 20–39% success rate with the same number of queries. Furthermore, we extend our evaluation to untargeted and targeted attacks on blackbox semantic segmentation models. Our method achieves 0.9–1.55% mIoU for untargeted and 69–95% pixel-wise success for targeted attacks. By comparison, the current state-of-the-art method [172] obtains 0.6–7.97% mIoU for untargeted attacks and does not report results for targeted attacks. To the best of our knowledge, our work is the first approach for targeted and query-based attacks for semantic segmentation.

Below we summarize main contributions of this work.

- We design a novel framework that can effectively attack blackbox dense prediction models based on an ensemble of surrogate models.
- We propose two simple yet highly effective ideas, namely weight balancing and weight optimization, with which we can achieve significantly better attack performance compared to existing methods.

- We extensively evaluate our method for targeted and untargeted attacks on object detection and semantic segmentation models and achieve state-of-the-art results.
- We demonstrate that our proposed method can generate a single perturbation that can fool multiple blackbox detection and segmentation models simultaneously.

3.2 Related work

Blackbox adversarial attacks. In the context of blackbox attacks, the attacker cannot access the model parameters or compute the gradient via backpropagation. Blackbox attack methods can be broadly divided into two groups: transfer-based [372, 373, 309, 281] and query-based attacks [92, 219, 459]. Transfer-based attacks rely on the assumption that surrogate models share similarities with the victim model, such that an adversarial example generated for the surrogate model can also fool the victim model. Query-based methods generate attacks by searching the adversarial examples space based on the feedback obtained from the victim model through queries. They can often achieve higher success rate but may require a large number of queries.

Ensemble-based attacks. Ensemble-based attacks leverage the idea of transfer attack and assume that if an adversarial example can fool multiple models simultaneously, the chances of fooling an unseen model are higher [309, 528, 131]. Recently, some methods have combined ensemble-based transfer attacks with limited feedback from the victim models to improve the overall success rate [177, 208, 445, 436, 284, 319]. These methods have mainly focused on classification models, and ensemble attacks on dense prediction tasks such as object detection and semantic segmentation are relatively less studied, especially for targeted attacks [500].

Attacks against object detectors and segmentation. Dense (pixel-level) prediction tasks such as object detection and semantic segmentation have higher task complexities[483] compared to classification tasks. Existing attacks on object detectors mainly focus on whitebox setting, although there are a few exceptions [490, 64]. A recent study [66] generates blackbox attacks on object detectors by using a surrogate ensemble and context-aware attack-based queries. Another approach [490] trains a generative model to generate transferable attacks. While some patch-based attacks [307, 406] are effective, the patches are easily noticeable. Recent works [173, 172] have investigated adversarial robustness for semantic segmentation and proposed a transferable untargeted attack using a single surrogate model. While most existing methods are based on a single surrogate model, we demonstrate that using multiple surrogates with weight balancing/search in the attack generation process, we can generate more effective adversarial examples for both untargeted and targeted scenarios, as well as for various types of dense prediction tasks.

3.3 Method

3.3.1 Preliminaries

We consider a per-instance attack scenario in which we generate adversarial perturbation δ for a given image x. To keep the perturbation imperceptible, we bound its ℓ_p norm as $\|\delta\|_p \leq \varepsilon$. In our experiments, we mainly use ℓ_∞ or max norm that limits the maximum level of perturbation. Our goal is to find δ such that the perturbed image, $x^\star = x + \delta$, can disrupt a victim image recognition system $f_\mathbf{v}$ to make wrong predictions. Suppose the original prediction for the clean image x is $y = f_\mathbf{v}(x)$. The attack goal is $f(x^\star) \neq y$ for untargeted attack, and $f(x^\star) = y^\star$ for targeted attack, where $y^\star$ is the desired output (e.g., label or bounding box or segmentation map).

For classification models, the label $y \in \mathbb{R}$ is a scalar. However, dense prediction models can have more complex output space. For object detection, the variable-length output $y \in \mathbb{R}^{K\times 6}$, where K is the number of detected objects, and each object label and position are encoded in a vector of length 6 that include the object category, bounding box coordinates, and confidence score. Some other tasks like keypoint detection and OCR are similar to object detection. For semantic segmentation, the prediction $y \in \mathbb{R}^{H\times W}$ is per-pixel classification, where H and W are the height and width of the input image, respectively. Depth and optical flow estimation tasks have similar output structure.

The adversarial loss functions for object detection and semantic segmentation can be defined using their respective training or prediction loss functions. Let us consider a whitebox model f and an input image x with output $y = f(x)$. For untargeted attack, we can search for the adversarial example $x^\star$ by solving the following maximization problem:

$$x^\star = \arg\max_x \ \mathcal{L}(f(x), y), \tag{3.1}$$

where $\mathcal{L}(f(x), y)$ represents the training loss of the model with input x and output y. For targeted attacks, with a target output $y^\star$, we solve the following minimization problem:

$$x^\star = \arg\min_x \ \mathcal{L}(f(x), y^\star). \tag{3.2}$$

Different from classification, which mostly use cross-entropy loss across different models, dense predictions have different loss functions for different models due to the complexity of the output space and diversity of the architectures. For example, two-stage object detector, including Faster RCNN [396], has losses for object classification, bounding box regression, and losses on the region proposal network (RPN). But for one-stage object detectors like YOLO [391, 393], they do

not have losses corresponding to RPN. Due to the large variability of the loss functions used in different dense prediction models, we use the corresponding training loss $\mathcal{L}$ for each model as the optimization loss to guide the backpropagation.

We employ PGD[326] to optimize the perturbation as

$$\delta^{t+1} = \Pi_\varepsilon \left(\delta^t - \lambda \, \mathbf{sign}(\nabla_\delta \mathcal{L}(f(x+\delta^t), y^\star)) \right), \tag{3.3}$$

for targeted attack and

$$\delta^{t+1} = \Pi_\varepsilon \left(\delta^t + \lambda \, \mathbf{sign}(\nabla_\delta \mathcal{L}(f(x+\delta^t), y)) \right), \tag{3.4}$$

for untargeted attack. Here t indicates the attack step, λ is the step size, and Π_ε projects the perturbation into a ℓ_p norm ball with radius ε. In the rest of the paper, we focus on targeted attacks without loss of generalization.

3.3.2 Ensemble-based attacks

In an ensemble-based transfer attack, we use an ensemble of N surrogate (whitebox) models: $\mathcal{F} = \{f_1, \ldots, f_N\}$ to generate perturbations to attack the victim model $f_{\mathbf{v}}$. Note that if the ensemble has a single model, then such an attack becomes a simple transfer attack with a single surrogate model. Let us denote the training loss function for ith model as $\mathcal{L}_i(f_i(x), y^*)$. A natural approach to combine the loss functions of all surrogate models is to compute an average or weighted average of the individual loss functions. For instance, we can generate the adversarial image by solving the following optimization problem:

$$x^\star(\boldsymbol{\alpha}) = \arg\min_x \sum_{i=1}^{N} \alpha_i \mathcal{L}_i(f_i(x), y^\star), \tag{3.5}$$

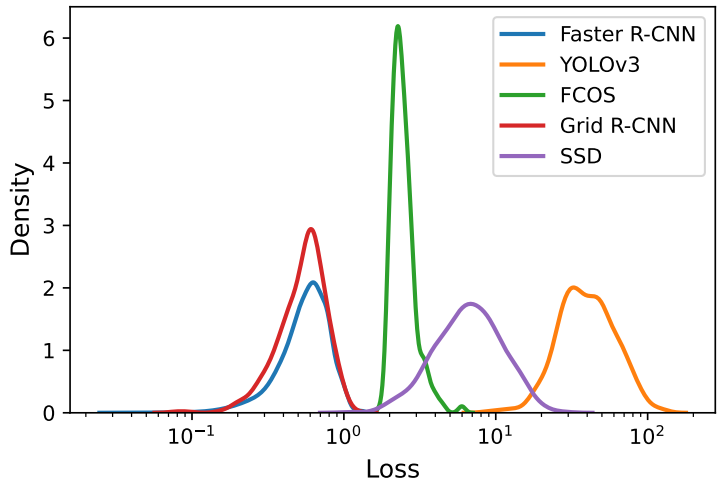


Figure 3.2: Distribution of losses for different object detection models. $\mathbb{P}(\mathcal{L}_i(f_i(x), y^\star))$. Calculated on 500 images from VOC dataset.

where $x^\star(\boldsymbol{\alpha})$ is a function of the weights of the ensemble $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_N\}$. One of our key observations is that the choice of weights plays a critical role in the transfer attack success rate of the ensemble models.

Weight balancing (victim model agnostic). In ensemble-based transfer attacks, we build on the intuition that if an adversarial example can fool all models simultaneously, it would potentially be more transferable to any unseen victim model. This concept has been empirically corroborated by numerous works [309, 131]. However, most attack methods have only been verified on classification models, all of which use the same cross-entropy loss and yield similar loss values. In contrast, the loss functions for object detectors in an ensemble can differ significantly and cover a large range of values (as shown in fig:obj-loss-imbalance). In such cases, models with large loss terms heavily influence the optimization procedure, reducing the attack success rate for models with small losses (see tab:obj-ablation). To overcome this issue, we propose a simple yet effective solution to balance the weights assigned to each model in the ensemble as follows. For each input image x and target

output $y^\star$, we adjust the weight for ith surrogate model loss as

$$\alpha_i = \frac{\sum_{i=1}^{N} \mathcal{L}_i(f_i(x), y^\star)}{N\mathcal{L}_i(f_i(x), y^\star)}. \tag{3.6}$$

The weights are adjusted in a whitebox setting as it allows us to measure the loss of each whitebox model accurately. The purpose of weight balancing is to ensure that all surrogate models can be successfully attacked, making the generated example more adversarial for blackbox victim models.

Weight optimization (victim model specific). Note that the weight normalization, as discussed above, is agnostic to the victim model. We further observe that such transfer-based attacks can be further improved by optimizing the weights of the ensemble according to the victim model, input image, and target output. In particular, we can change the individual α_i to create the perturbations that reduce the victim model loss $\mathcal{L}_{\mathbf{v}}$. To achieve this goal, we need to solve the following optimization problem with respect to $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^\star = \arg\min_{\boldsymbol{\alpha}} \ \mathcal{L}_{\mathbf{v}}(f_{\mathbf{v}}(x^\star(\boldsymbol{\alpha})), y^\star). \tag{3.7}$$

The optimization problem in (3.7) is a nested optimization that we can solve as an alternating minimization routine.

Step 0. Given input x, output $y^\star$, and surrogate ensemble $\mathcal{F}$, we initialize $\boldsymbol{\alpha}$ using (3.6).

Step 1. Solve (3.5) to generate an adversarial example $x^\star(\boldsymbol{\alpha})$.

Step 2. Test the victim model. Stop if attack is successful; otherwise, change one of the α_i and repeat Step 1.

In our experiments, we update the α_i in a cyclic manner (one coordinate at a time) as $\alpha_i \pm \gamma$ in **Step 2**, where γ denotes a step size. In every round, we select the value of α_i that provides smallest value of the victim loss. We count the number of queries as the number of times we test the generated adversarial example on the victim model and denote it as Q in our experiments.

3.4 Experiments

To evaluate the effectiveness of our method, we performed extensive experiments on attacking various object detection and semantic segmentation models. We first show that the attacks generated by a single surrogate model fail to transfer to arbitrary victim models. Then we show that the attack transfer rate can be increased by using an ensemble with weight balancing. Additional optimization of the weights surrogates for each victim model can further improve the attack performance. Finally, we show that we can generate single perturbations to fool object detectors and semantic segmentation models simultaneously.

3.4.1 Experiment setup

Object detection

Models and datasets. We utilize `MMDetection` [90] toolbox to select various model architectures and weights pre-trained on COCO 2017 dataset [302]. To construct the surrogate ensemble, we start with two widely used models, `Faster R-CNN` [396] and `YOLO` [391, 393], and expand the ensemble by appending models with different architectures, including {`FCOS` [450], `Grid R-CNN` [323], `SSD` [305]}. We select different victim models, including {`RetinaNet` [301], `Libra R-CNN` [370], `FoveaBox` [257], `FreeAnchor` [539], `DETR` [76]}. We evaluate attack performance on COCO 2017 [302] and Pascal VOC 2007 [140] datasets. Since the models from this repository are trained on COCO, which contains 80 object categories (a superset of VOC dataset's 20 categories), while testing on VOC dataset, we only return the objects that exist in VOC. We follow the setup in [66] and randomly select 500 images containing multiple (2–6) objects from VOC 2007 test and COCO 2017 validation sets.

Evaluation metrics. We mainly focus on targeted attacks for object detection since they are more challenging than untargeted or vanishing attacks. We measure the performance of the attack using attack success rate (ASR), which equals the number of successfully attacked images over the total number of attacks. We follow the setting in [66], where if the target label is detected within the victim object region with IOU > 0.3, the attack is determined a success.

Perturbation and query budget. We tested different perturbation levels with $\ell_\infty = \{10, 20, 30\}$ out of 255. We use at most 10 queries for attacking object detectors, and we show the trends of how ASR increases with the number of queries. To align with [66] that uses 5 attack plans, we set the maximum query budget to $Q = 5$ in tab:obj-ablation.

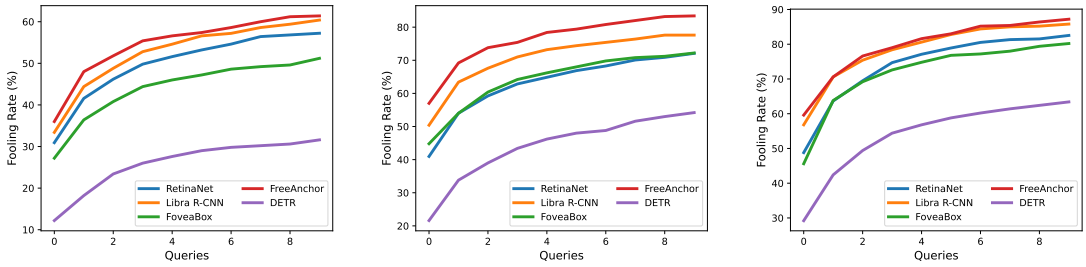
Comparing methods. We compare with [66], which is a state-of-the-art transfer-based approach that leverages context information to design attack plans to iteratively attack the victim object. The method generates different perturbations by iterating over a set of predefined attacks, and the total number of queries is the number of attempted attacks. BASES [64] is a recent work on ensemble-based blackbox attacks, which mainly focused on classification tasks and did not consider the loss distributions of different surrogate models. In our experiments, the ensemble with weight optimization and without balancing is equivalent to BASES [64].

Semantic segmentation

Models and datasets. We use `MMSegmentation` [109] toolbox to select different model architectures and weights pre-trained on Cityscapes [111] ($x \in \mathbb{R}^{512\times1024\times3}$) and Pascal VOC ($x \in \mathbb{R}^{512\times512\times3}$) datasets. We select `PSPNet` [542] and `DeepLabV3` [91] with `ResNet50` and `ResNet101` [187] backbones as our blackbox victim models. For the surrogate ensemble, we start with the pri-

Table 3.1: Targeted attack success rate (%) of different methods at different perturbation budgets on VOC dataset. For each perturbation level, the first 4 rows correspond to different settings of our attacks, *i.e*. with (✓) or without (✗) weight balancing and weight optimization. We show comparison with context-aware attack [66], the state-of-the-art method for query-based blackbox attacks.

Perturbation	Weight	Weight	**Surrogate Ensemble**		**Blackbox Victim Models** (ASR ↑)				
Budget	Balancing	Optimization	FRCNN	YOLOv3	Retina	Libra	Fovea	Free	DETR
	✗	✗	27.9	91.5	11.6	9.2	9.0	13.4	5.6
	✗	✓	61.4	**99.4**	24.3	28.0	22.4	31.0	15.4
$\ell_\infty = 10$	✓	✗	71.1	85.7	30.9	33.4	27.2	36.0	12.2
	✓	✓	**86.0**	96.9	**53.2**	**56.6**	**47.2**	**57.4**	**29.0**
	Context-aware Attack [66]		55.8	75.6	22.6	20.4	33.6	39.2	20.2
	✗	✗	40.1	92.2	16.9	20.4	15.4	23.2	9.7
	✗	✓	77.7	**99.8**	41.0	45.4	37.8	47.0	22.5
$\ell_\infty = 20$	✓	✗	82.7	89.8	41.0	50.4	44.8	57.0	21.6
	✓	✓	**94.6**	98.0	**66.9**	**74.4**	**68.0**	**79.4**	**48.0**
	Context-aware Attack [66]		78.6	87.2	35.2	38.4	51.6	56.6	34.0
	✗	✗	43.4	91.1	17.1	22.6	17.4	27.2	11.4
	✗	✓	82.7	**99.6**	47.2	54.8	47.0	57.4	33.4
$\ell_\infty = 30$	✓	✗	85.3	90.2	48.8	56.8	45.6	59.6	29.2
	✓	✓	**96.0**	98.1	**78.9**	**82.8**	**76.8**	**83.0**	**58.8**
	Context-aware Attack [66]		80.6	88.0	42.0	44.2	56.8	63.6	40.2



(a) $\ell_\infty \leq 10$ **(b)** $\ell_\infty \leq 20$ **(c)** $\ell_\infty \leq 30$

Figure 3.3: Attack success rate (or fooling rate) vs number of queries (Q). The maximum value of Q is set to 10 for these results.

mary semantic segmentation model `FCN` [317], and expand the ensemble with {`UPerNet` [504], `PSANet` [543], `GCNet` [75], `ANN` [555], `EncNet` [532]}. All models are built on `ResNet50` [187] backbone trained with the cross-entropy loss. The loss values across all surrogate models have similar range; therefore, the effect of weight balancing for semantic segmentation is not as significant as it is for object detection. We use validation datasets from Cityscapes [111] and Pascal VOC 2012, which contains 500 and 1499 images with 19 and 21 classes, respectively.

Evaluation metrics. We use different metrics for untargeted and targeted attack performance evaluation. In untargeted experiments, the attack performance is evaluated using the mIoU score (in percentage %), the lower mIoU score the better attack performance. For targeted experiments, we report the pixel success ratio (PSR), which indicates the percentage of pixels successfully assigned the desired label in the target region, the higher the better attack performance.

Perturbation and query budget. We use the perturbation budget $\ell_\infty \leq 8$ out of 255 and query budget $Q = 20$.

Table 3.2: Targeted ASR (%) for blackbox victim models and whitebox surrogate models with different ensemble sizes (N). On VOC dataset, $\ell_\infty \leq 10$.

N	**Surrogate Ensemble**					**Blackbox Victim Models** (ASR ↑)				
	FRCNN	YOLOv3	FCOS	Grid R-CNN	SSD	Retina	Libra	Fovea	Free	DETR
1	74.7	-	-	-	-	31.3	31.2	29.8	40.4	10.6
2	86.0	96.9	-	-	-	53.2	56.6	47.2	57.4	29.0
3	87.9	96.1	74.2	-	-	63.1	62.0	57.3	66.6	38.0
4	89.6	94.7	75.2	87.9	-	68.7	**71.0**	67.6	74.4	49.6
5	89.7	91.8	73.5	86.1	82.4	**68.9**	70.2	**68.4**	**77.6**	**53.2**

Comparing methods. We compare with dynamic scale (DS) attack [172] which is the most recent method that achieves the highest attack transfer rate on semantic segmentation untargeted attacks.

3.4.2 Attacks against object detection

Following settings in [66], we randomly select one object from the output of victim model as the victim object and perturb it into a target object that does not exist in the original detection. This approach rules out the possibility of mis-counting existing objects as the target object.

We report our main results in tab:obj-ablation. The baseline method uses a surrogate ensemble without weight balancing and models are assigned weight of 1. Such a baseline method is same a transfer-based method and results in highly imbalanced success rate for different surrogate models. For instance, at $\ell_\infty \leq 10$, the success rate for `YOLOv3` is above 90% while the success rate for `Faster R-CNN` is less than 30%. Low success rate on surrogate side translates to low success

Table 3.3: Untargeted attack mIoU scores (%) of ensemble sizes $N = 2, 4, 6$ on Cityscapes dataset. We compare $Q = 0$ (i.e. direct transfer attack) with $Q = 20$ ensemble attack performance. DS uses DeepLabV3-Res50 (DL3-50) as the surrogate model for attack generation; thus the DS on DL3-50 is a whitebox attack. While our method used an ensemble that does not include any victim models for attack generation, we still achieved comparable mIoU scores to DS on DL3-50. Blue numbers represent whitebox attacks.

Method	**Whitebox Surrogate**	**Blackbox Victim Models** (mIoU ↓)			
		PSPNet-Res50	PSPNet-Res101	DeepLabV3-Res50	DeepLabV3-Res101
Clean Images	-	77.92	78.28	79.12	77.12
Baseline	PSPNet-Res50	3.43	24.18	5.05	25.74
	DeepLabV3-Res50	4.76	21.72	3.92	22.23
DS[172]	PSPNet-Res50	0.82	8.04	1.36	9.00
	DeepLabV3-Res50	1.23	7.97	0.61	7.11
Ours ($Q = 0$)	$N = 2$	5.07	8.32	5.19	8.74
	$N = 4$	4.33	6.26	4.32	6.33
	$N = 6$	3.62	4.91	4.02	4.84
Ours ($Q = 20$)	$N = 2$	1.38	2.88	1.15	3.50
	$N = 4$	**0.79**	2.04	**0.73**	1.80
	$N = 6$	0.90	**1.55**	0.94	**1.09**

rate on blackbox victim side. The main reason for such imbalance is that the loss of different object detectors can be highly unbalanced (e.g., the loss value for `YOLOv3` is nearly 60× larger than the loss of `Faster RCNN` for targeted attacks, *c.f.* fig:obj-loss-imbalance). With weight balancing, the success rate increases for surrogate and blackbox victim models. The success rate is further increased on surrogate and victim blackbox models if we optimize the weights, same as BASES [64]. Our method (with weight balancing and optimization) achieves a significantly higher ASR

Table 3.4: Targeted attack performance on Cityscapes as pixel success rate (higher the better). The attack performance increases as we increase ensemble size (N) and number of queries for weight optimization (Q). $N = 1$ has zero query. We note PSPNet-Res50 as PSP-r50, and DeepLabV3-Res50 as DL3-r50, similar abbreviations apply to Res101.

Q	N	**Blackbox Victim Models** (PSR ↑)			
		PSP-r50	PSP-r101	DL3-r50	DL3-r101
0	1	39.15	10.21	35.02	7.58
	2	52.15	12.28	47.99	10.59
	3	43.17	11.34	42.10	9.87
	4	51.44	26.13	49.14	17.42
	5	52.24	23.88	51.75	16.08
20	2	83.97	51.80	82.70	46.95
	3	88.88	64.63	85.55	60.88
	4	91.51	64.28	87.19	63.88
	5	**92.91**	**69.09**	**88.95**	**69.65**

compared to context-aware attack across different datasets and different perturbation budgets. On average, our ASR on blackbox victim models is over $4\times$ better than baseline method and over $1.5\times$ better than context-aware attack. On whitebox surrogate models, weight balancing and optimization also achieves the highest ASR. Context-aware attack fixes weight ratio for surrogate models, $\alpha_{\text{FRCNN}}\alpha_{\text{YOLO}} = 4$, which is sub-optimal according to our analyses. Even though it achieves much higher performance than baseline, it still largely under-performs our method. Similar trend is observed for COCO dataset (see tab:obj-ablation-coco).

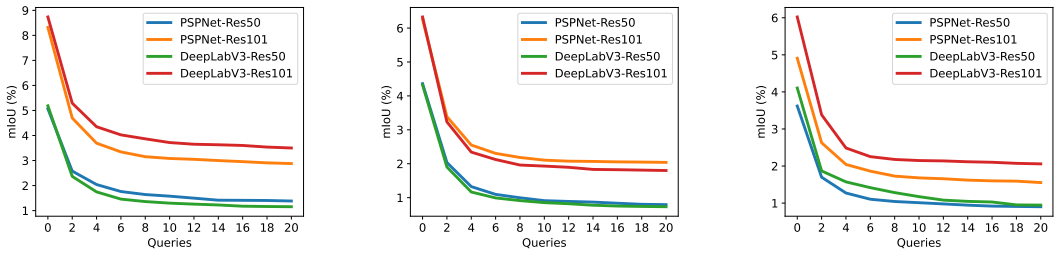
fig:obj-trend shows the effect of the number of queries on the ASR that gradually improves as we optimize the weights. We observe the largest increase in the first two steps and then the improvement plateaus as $Q \rightarrow 10$.

We also conducted an experiment to test our method with varying ensemble sizes. The results for $\ell_\infty \leq 10$, $Q = 5$ are presented in tab:obj-ensemble-size. As we increase the number of models in the ensemble from $N = 1$ to $N = 5$, we observe an increased ASR on all blackbox victim models.

3.4.3 Attacks against semantic segmentation

We evaluate the effectiveness of our attack on semantic segmentation in both untargeted and targeted settings. For the sake of consistency and a fair comparison, we adopt adversarial attack settings in DS attack [172].

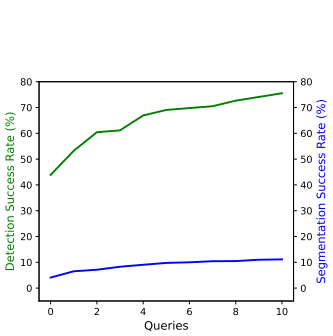
Untargeted attacks. We generate adversarial attacks using different ensemble sizes and report mIoU scores on Cityscapes in tab:seg-untar-cs and fig:seg-trend (and Pascal VOC in supplementary material). In the untargeted setting, semantic segmentation models are attacked to maximize



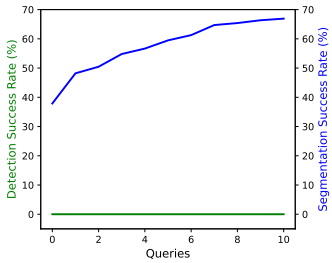
(a) $N = 2$ **(b)** $N = 4$ **(c)** $N = 6$

Figure 3.4: mIoU vs number of queries (Q) for different ensemble sizes (N).

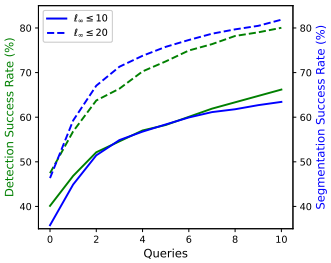
the loss between clean and modified annotation; hence, the lower mIoU implies better attack performance. All of the victim models achieve high performance on clean images. The baseline method (direct transfer attack with one surrogate model using PGD) performs well in the whitebox setting but suffers when the victim uses another backbone. For example, the attacks generated on `PSPNet-Res50` achieves 3.43% mIoU on `PSPNet-Res50` but only attains 24.18% mIoU on `PSPNet-Res101`. DS attack achieves better results than the baseline method but still suffers from cross-backbone transfers. On the other hand, our method, without weight optimization (*i.e.*, $Q = 0$) and using a surrogate ensemble of $N = 2$ models, can achieve results comparable to DS attack, particularly for attacks on Res101 models. As we increase the number of surrogate models to 4 or 6, our attack performance further improves. Furthermore, when we apply weight optimization (*e.g.*, $Q = 20$), the attack improves by updating the weights of the surrogate models, allowing us to outperform DS attack for all victim models. fig:seg-trend shows how the mIoU changes with the number of queries. We observe that the mIoU gradually reduces as we query the victim model and optimize the weights. The largest decrease happens in the first 3–4 steps and then the reduction plateaus as $Q \to 20$.



(a) Attacks generated for object detection only, $\ell_\infty \leq 10$



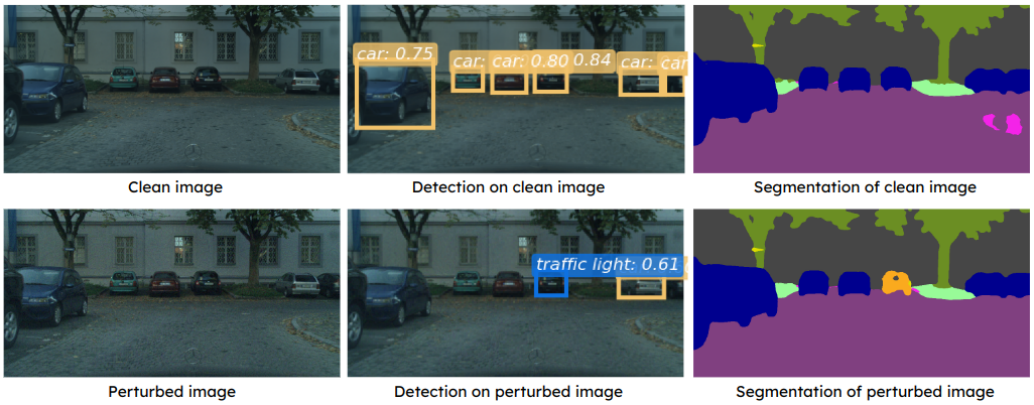
(b) Attacks generated for semantic segmentation only, $\ell_\infty \leq 10$



(c) Attacks generated jointly for object detection and semantic segmentation, $\ell_\infty \leq \{10, 20\}$

Figure 3.5: Comparison between task-specific attacks and joint attack performance on blackbox object detector (`RetinaNet`) and segmentation model (`PSPNet`). Green curves denote attack success rate for object detectors, and blue curves denote pixel success rate for semantic segmentation. (a) Attacks generated with an object detector surrogate do not transfer for semantic segmentation. (b) Attacks generated with semantic segmentation models surrogate do not transfer for object detectors. (c) Attacks generated by a surrogate of object detectors and semantic segmentation (along with weight balancing and optimization) provide successful attacks for blackbox object detectors and semantic segmentation models.

Targeted attack. To evaluate our method in a more challenging setting, we consider a targeted attack scenario, where instead of changing every pixel in the segmentation to some arbitrary label, we focus on attacking a dominant class (*i.e.*, the class occupying the largest area) in the scene to its least likely class $y^\star$. For each clean image, we first select a region with the dominant class y (*e.g.*, "road" or "building" for most of the Cityscapes images. See fig:seg-tgt-explain as an example).



(a) Our method can generate attacks to fool multiple blackbox object detector and blackbox semantic segmentation models jointly. First row shows a clean image from Cityscapes dataset, detection with `RetinaNet` and segmentation with `PSPNet`. Second row shows the perturbed image using ensemble surrogates {`Faster RCNN`, `YOLOv3`, `FCN`, `UPerNet`}, and detection and segmentation results on the perturbed image. We generate perturbation to map the Car in the middle to Traffic Light. Image id: `lindau_000026_000019`

(b) Color encoding for segmentation maps in CityScapes dataset

Figure 3.6: Visual adversarial examples of our method that generates successful attacks to fool a blackbox object detector and a blackbox semantic segmentation model using a single perturbed image.

Then based on the least-likely class of each pixel in that region, we select the class that appears most frequently as the target label $y^\star$ of the entire region. We use PSR as our evaluation metric, which represents the percentage of pixels in the selected region that are successfully assigned to $y^\star$. The higher percentage indicates more pixels are successfully attacked to the desired class, which

indicates better attack performance. Our targeted attack results are reported in tab:seg-targeted. Results show that as we increase the number of surrogate models (N), the ASR improves for most instances without any weight optimization step (i.e., $Q = 0$). If we perform weight optimization for $Q = 20$ steps, then the success rate increases for all the models. For instance, with $N = 4$, the ASR for Res101 models increases from 17–26% to 63-64%.

3.4.4 Joint attack for multiple models and tasks

We first show that generally adversarial examples generated for object detection do not transfer to semantic segmentation, and vice versa. Then we show that we can generate single perturbations to fool object detectors and semantic segmentation models simultaneously, by using a surrogate ensemble including both detection and segmentation models. We choose targeted attacks in our experiments because they are more challenging than untargeted attacks.

Experiment setup. On the blackbox (victim) side, we tested `RetinaNet` as the victim object detector and `PSPNet-Res50` as the victim semantic segmentation model. On the whitebox (surrogate) side, we used `Faster RCNN, YOLOv3` as the surrogate object detectors and `FCN, UPerNet` as the surrogate semantic segmentation models. We performed targeted attacks on 500 test images selected from the validation set of CityScapes dataset.

Results. We present the ASRs for task-specific and joint attacks in fig:obj-joint-curves. Green curves denote ASR for object detectors, and blue curves denote PSR for semantic segmentation. fig:sub-joint-det presents the results when we generate attacks using an object detector surrogate ensemble. Note that success rate for victim object detector (`RetinaNet`) increases as we optimize the weights but the success rate for the semantic segmentation model (`PSPNet`) remains small. Similarly, fig:sub-joint-seg presents the results when we generate attacks using a segmentation surrogate

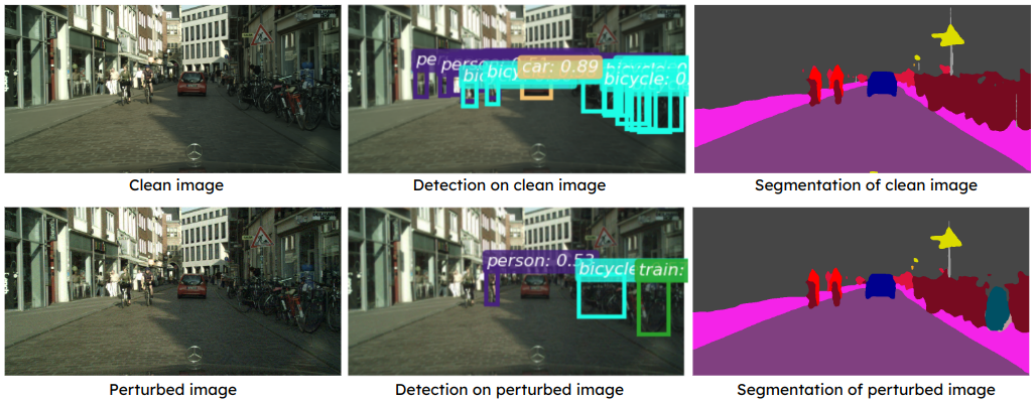
ensemble. The success rate for the victim semantic segmentation model increases, but the success rate for the object dector remains close to zero. fig:sub-joint-all presents the results when we perform a joint attack using an ensemble that consists of both object detectors and segmentation models. The blackbox ASR is high on both detection and segmentation fig:sub-joint-all, and the attack performance improves as we update the weights of the surrogate models. In fig:sub-joint-all, we show the results for different perturbation budgets, with $\ell_\infty \leq 10$, the success rates on detection and segmentation are between $60\% - 70\%$, which are close to in-domain detection attacks in fig:sub-joint-det and in-domain segmentation attacks fig:sub-joint-seg. When we increase the perturbation to $\ell_\infty \leq 20$, the success rate for both detection and segmentation can surpass 80%.

Visualization of adversarial examples. In this example, our goal is to perturb the car in the middle to a traffic light. We assign the target label for car region to traffic light. fig:obj-joint-attack shows the results where a single adversarial image generated by the surrogate model can successfully fool the blackbox models `RetinaNet` and `PSPNet`.

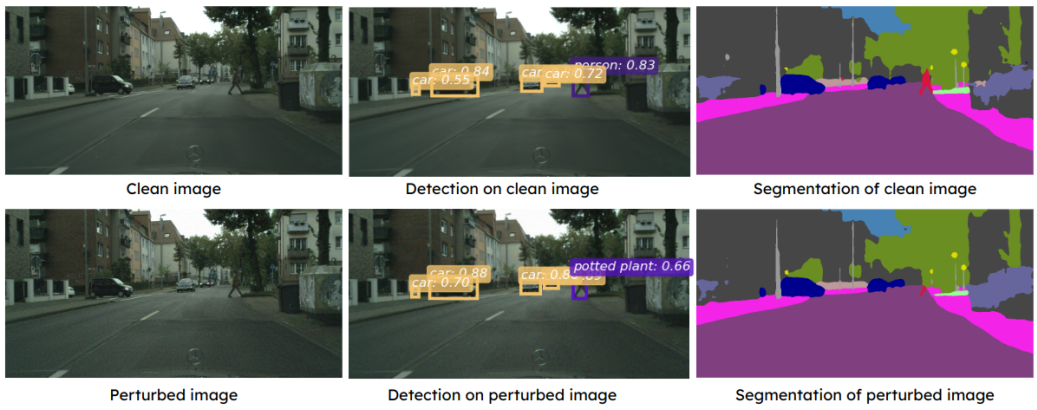
3.4.5 Joint attack for multiple blackbox models

In this section, we provide additional visualization results for joint (targeted) blackbox attacks against object detection and semantic segmentation models.

More Visualization of adversarial examples. We visualize some adversarial examples in fig:obj-joint-examples. In fig:joint-example1, we show an example where our method generates a single perturbed image to map the bicycle on the right-hand-side to train. In object detection results we see the label for the Bicycle bounding box has been changed to Train, and for the segmentation map, the corresponding region has changed to teal color encoding for Train as well. In fig:joint-example2,



(a) Generate perturbation to map the Bicycle on the right-hand-side to Train. Image id: `munster_000140_000019`



(b) Generate perturbation to map the Pedestrian to Potted Plant. Image id: `munster_000006_000019`

(c) Color encoding for segmentation maps in CityScapes dataset

Figure 3.7: Visual adversarial examples of our method that generates successful attacks to fool a blackbox object detector and a blackbox semantic segmentation model using a single perturbed image.

Table 3.5: Targeted attack success rate (%) for different methods on COCO dataset. Similar setting as in tab:obj-ablation.

Perturbation	Weight	Weight	**Surrogate Ensemble**		**Blackbox Victim Models** (ASR ↑)				
Budget	Balancing	Optimization	FRCNN	YOLOv3	RetinaNet	Libra	Fovea	Free	DETR
	✗	✗	19.6	79.7	4.6	5.0	4.4	6.6	2.6
	✗	✓	49.8	**97.8**	13.5	16.2	14.4	22.6	8.4
$\ell_\infty = 10$	✓	✗	57.2	65.3	16.2	17.6	16.6	24.0	5.4
	✓	✓	**78.0**	86.1	**31.7**	**32.0**	**32.3**	**41.6**	**15.4**
	Context-aware Attack [66]		41.2	54.4	12.0	11.2	18.6	25.0	10.8
	✗	✗	25.8	82.2	8.9	9.8	8.4	13.2	5.6
	✗	✓	62.4	**98.2**	23.0	32.2	22.4	32.2	13.2
$\ell_\infty = 20$	✓	✗	68.8	75.8	25.5	28.0	27.1	38.0	13.8
	✓	✓	**88.9**	94.5	**48.5**	**53.8**	**49.5**	**65.6**	**31.0**
	Context-aware Attack [66]		64.4	70.0	20.8	22.2	35.4	40.8	20.0
	✗	✗	29.0	82.2	8.8	9.4	13.3	14.6	6.4
	✗	✓	69.0	**99.3**	27.9	34.6	31.2	43.6	17.6
$\ell_\infty = 30$	✓	✗	72.7	78.6	32.5	33.8	34.1	41.6	14.8
	✓	✓	**91.7**	95.5	**57.6**	**64.4**	**58.3**	**71.2**	**36.6**
	Context-aware Attack [66]		68.6	75.4	27.2	27.2	39.2	46.2	21.2

the generated perturbed image maps the car in the middle to traffic light. Note that the bounding box for the Car in the middle changed to Traffic Light for the object detector and the same area in the semantic segmentation map changed the color to orange (corresponding to Traffic Light label).

3.4.6 Attacks against object detection

Attacks using different surrogate models. In our previous experiments (tab:obj-ablation and tab:obj-ablation-coco), we follow the model selection in [66] for a fair comparison. We can easily replace

the surrogate models with different ones and expect its effectiveness across different settings. For example, we can replace YOLOv3 with Deformable DETR (denoted as Deform) and get similar results, as shown in tab:obj-replaceYOLO below. The experiment setup and victim models are same as reported in tab:obj-replaceYOLO for $\ell_\infty = 20$.

Table 3.6: Replacing YOLOv3 with Deformable DETR. Correspond to tab:obj-ablation, perturbation budget $\ell_\infty = 20$.

Weight	Weight	**Surrogate Ensemble**		**Blackbox Victim Models** (ASR ↑)				
Balancing	Optimization	FRCNN	Deform	Retina	Libra	Fovea	Free	DETR
✗	✗	8.5	69.5	10.6	4.0	8.0	10.5	12.0
✗	✓	34.6	94.3	36.7	25.0	33.5	53.0	38.5
✓	✗	68.0	80.5	47.2	38.5	37.5	57.5	26.5
✓	✓	**88.1**	**95.0**	**74.9**	**70.5**	**73.0**	**84.0**	**56.0**
ZQA [63]		88.2	-	44.0	51.4	53.4	-	-

Comparisons with zero-query attacks. Zero-Query attack (ZQA) [63] does not rely on any feedback from the victim. It assesses the attack success probability on the surrogate model before launching a single and most promising attack against the victim. Due to these differences in problem setting, we do not directly compare with this method in the main paper. Here we compare the numbers reported from corresponding manuscripts in tab:obj-replaceYOLO. ZQA uses a single surrogate model without any feedback from the victim model. It performs worse than the few-query attacks [66] with 3–5 queries, and our method clearly outperforms both of them.

Comparison with conventional query-based attacks. Existing query-based methods, including GARSDC [297] and PRFA [298], require thousands of queries (which is prohibitive) and they are

only applicable for untargeted attacks. Furthermore, their perturbations are clearly visible, see Fig. 5 in [297], while our perturbations remain imperceptible. For these key differences, we did not include their comparison in the main paper, but here we provide a mAP score comparison with them. We use 5 surrogate models from tab:obj-ensemble-size and perform vanishing attacks on ATSS [538] model, we show in tab:compare-conventional that our method can achieve a near-zero mAP within just a few queries (Q).

Table 3.7: Comparison with conventional query-based attacks.

Method	ATSS[538]	
	mAP ↓	Q ↓
Clean	0.54	N/A
PRFA [298]	0.20	3500
GARSDC [297]	0.04	1837
Ours	**0.00**	**10**

3.4.7 Attacks against semantic segmentation

Attacks on Pascal VOC dataset. We generate adversarial attacks using different sizes of ensemble and report mIoU scores on the Pascal VOC dataset in tab:seg-untar-voc. Similar to the results on the Cityscapes dataset in Tab. tab:seg-untar-cs, as we increase the number of surrogate models from 2 to 6, the attack performance improves (indicated by smaller mIoU scores). Attack performance of our method further improves with weight optimization (with $Q = 20$). These results show that by adjusting the weights of the surrogate ensemble, we can improve the attack performance. Our attack method with $N = 6$ surrogate models provides 27–29% improvement in mIoU scores com-

pared to DS attack for the victim models PSPNet-Res50 and DeepLabV3-Res50. Note that DS attack uses these two models as the whitebox surrogates as well victim models. In contrast, we keep all four victim models PSPNet-Res50, DeepLabV3-Res50, PSPNet-Res101, DeepLabV3-Res101 out of our ensemble. Our surrogate ensemble consists of FCN, UPerNet, PSANet, GCNe, ANN, EncNet with ResNet50 backbones, which reflects a more realistic setting where the victim blackbox model is different from any of the surrogate models.

Effect of backbones on attack performance. We note that for VOC dataset results in tab:seg-untar-voc, our method provides high attack success for blackbox victim models with ResNet50 backbone. However, the attack performance on victim models with ResNet101 backbone degrades (as reflected by large mIoU values). To further demonstrate the effectiveness of our attack, we replace the backbones of the surrogate models with the ResNet101 backbones while keeping the rest of model architectures same as the original ensemble. Results reported in tab:seg-untar-voc-r101 show that if we replace surrogate models with ResNet101 backbones (same backbone as the victim blackbox models), then our attack method provides significantly better results.

Attack performance on different backbones. We performed additional experiments using FCN and PSPNet methods and MobileNetV2 and ResNeSt (denoted as -mv2 and -s101 in tab:seg-backbones) backbones for victim models. The attack setting corresponds to Tab. tab:seg-targeted. Due to the great difference in backbones across surrogate and victim, the attack performance drops. Nevertheless, the attack performance improves significantly as we increase the ensemble size and optimize ensemble weights. Results are reported in tab:seg-backbones.

Attack performance on surrogate models. For the sake of completeness, we also report attack performance on the whitebox surrogate models for both untargeted and targeted attacks in tab:seg-

Table 3.8: Semantic segmentation targeted pixel success ratio (PSR) (%) for blackbox victim models with different backbones.

Q	N	**Blackbox Victim Models** (PSR ↑)			
		FCN-mv2	FCN-s101	PSP-mv2	PSP-s101
0	1	33.26	1.01	3.96	2.71
	3	30.39	1.39	5.82	6.94
	5	38.92	3.12	7.84	8.32
20	3	50.31	22.79	24.09	54.06
	5	**53.09**	**34.57**	**30.20**	**60.43**

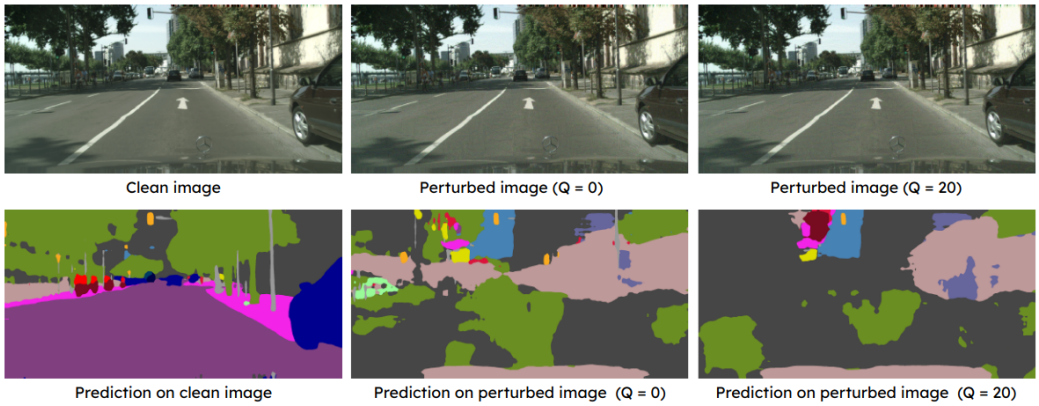
ensemble-size-untar and tab:seg-ensemble-size-tar. We observe that as we increase the number of models in the ensemble from $N = 1$ to $N = 5$, we can achieve better attack performance on all the whitebox and blackbox victim models we tested. Attacks that are successful on blackbox victim models are almost always successful on all surrogate models.

Visualization of adversarial examples. We present some visual examples of untargeted attacks in fig:seg-untar-examples and targeted attacks in fig:seg-tar-examples. We observe that the attacks generated by surrogate model do not transfer to the victim model for untargeted or targeted cases (i.e., $Q = 0$). The attacks generated after weight optimization (i.e., $Q = 20$) succeed for untargeted and targeted attacks. Our targeted attack setup is visually explained in fig:seg-tgt-explain. Instead of mapping every pixel prediction to an arbitrary target label, we focus on attacking a single object y in the original prediction (e.g.“road” in fig:tgt-original with white bounding-box). We select the

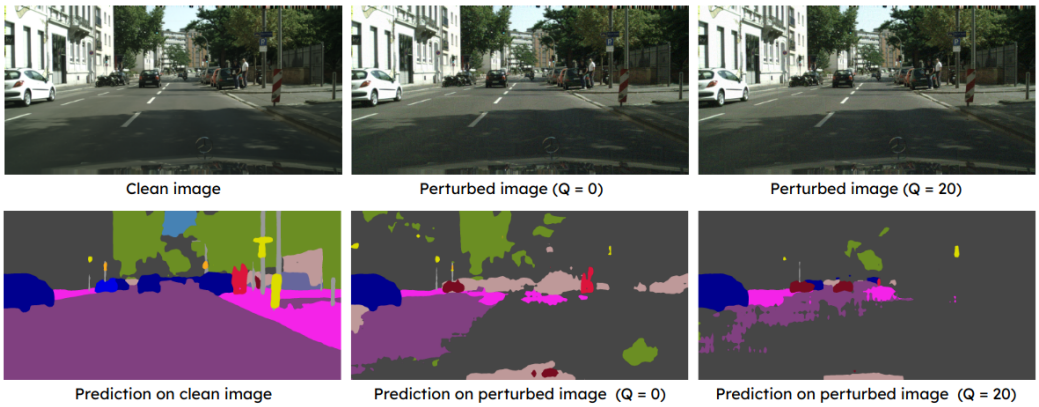
Table 3.9: mIoU scores (%) for untargeted attacks on semantic segmentation models with Pascal VOC dataset. The lower value indicates better attack performance. Surrogate of ensemble sizes $N = 2, 4, 6$. We compare $Q = 0$ (i.e. direct transfer attack) with $Q = 20$ ensemble attack performance. Results show enabling the ensemble query introduced attack performance increments. Blue numbers represent whitebox attacks.

Method	**Whitebox Surrogate**	**Blackbox Victim Models** (mIoU ↓)			
		PSPNet-Res50	PSPNet-Res101	DeepLabV3-Res50	DeepLabV3-Res101
Clean Images	-	76.78	78.47	76.17	78.70
Baseline	PSPNet-Res50	5.09	37.06	6.57	38.98
	DeepLabV3-Res50	3.63	22.01	3.14	22.58
DS	PSPNet-Res50	2.07	16.10	2.56	18.57
	DeepLabV3-Res50	2.31	**12.32**	2.15	**13.64**
Ours ($Q = 0$)	$N = 2$	14.33	35.47	12.31	35.31
	$N = 4$	8.74	29.41	7.92	28.01
	$N = 6$	7.28	24.28	6.75	24.63
Ours ($Q = 20$)	$N = 2$	5.56	27.49	4.43	28.46
	$N = 4$	2.23	22.24	2.09	20.34
	$N = 6$	**1.69**	18.07	**1.53**	17.61

target label $y^\star$ as the class that appears most frequently as the least-likely label of the pixels in the selected region. For example, fig:tgt-ll shows class "building" in grey color as the least likely class in the target region. Finally, we generate attack to replace the entire selected region in the original prediction to its target label (fig:tgt-target).



(a) Generate perturbation to maximize prediction error. We use the $N = 2$ surrogate ensemble to generate this attack against blackbox victim `PSPNet-Res50`. Image id: `frankfurt_000001_005703`



(b) Generate perturbation to maximize prediction error. We use the $N = 2$ surrogate ensemble to generate this attack against blackbox victim `DeepLabV3-Res101`. Image id: `frankfurt_000000_022797`

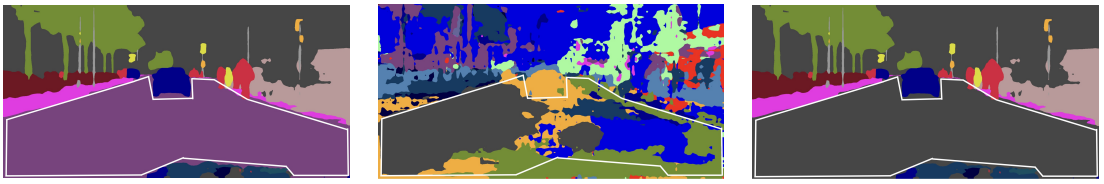
Figure 3.8: Visual adversarial examples of our method for untargeted attacks to fool a blackbox semantic segmentation model.

Table 3.10: mIoU scores (%) for untargeted attacks on semantic segmentation models with Pascal VOC dataset. The lower value indicates better attack performance. Surrogate of ensemble sizes $N = 1$ to 6. We compare the performance of `ResNet50` and `ResNet101` backbones in the ensemble. The attack performance on `ResNet101` backbone victim models increases if we use the surrogate models with `ResNet101` backbone. Note there is no weight optimization for $N = 1$.

Q	N	**Blackbox Victim: PSPNet-Res101** (mIoU ↓)		**Blackbox Victim: DeeplabV3-Res101** (mIoU ↓)	
		Backbone: Res50	Backbone: Res101	Backbone: Res50	Backbone: Res101
0	1	38.51	24.76	38.66	25.98
	2	35.47	21.50	35.31	21.54
	3	31.95	17.65	32.39	18.02
	4	29.41	14.53	28.01	14.32
	5	25.82	13.67	24.79	12.28
	6	24.28	12.49	24.63	12.35
20	2	27.49	8.78	28.46	8.80
	3	24.80	5.15	22.55	5.69
	4	22.24	5.49	20.34	4.49
	5	19.62	**3.27**	18.31	**3.13**
	6	**18.07**	4.04	**17.61**	3.32

Table 3.11: Semantic segmentation untargeted attack mIoU scores (%) for blackbox victim models and whitebox surrogate models with different ensemble sizes (N). The lower value indicates better attack performance. Experiment with CityScapes dataset, $\ell_\infty \leq 8$. PSP-r50, PSP-r101, DL3-r50, DL3-r101 stands for `PSPNet` and `DeepLabV3` built on `ResNet50`, `ResNet101` backbone respectively.

N	**Surrogate Ensemble**						**Blackbox Victim Models** (mIoU ↓)			
	FCN	UPerNet	PSANet	GCNet	ANN	EncNet	PSP-r50	PSP-r101	DL3-r50	DL3-r101
1	2.42	-	-	-	-	-	2.68	6.92	5.16	10.13
2	1.28	1.06	-	-	-	-	1.38	2.88	1.15	3.50
3	1.45	1.06	1.05	-	-	-	1.13	2.39	0.95	2.67
4	1.25	0.97	0.87	0.91	-	-	**0.79**	2.04	**0.73**	1.80
5	1.18	0.96	0.93	0.91	1.14	-	0.78	1.69	0.89	2.09
6	1.26	1.08	1.08	1.05	1.25	1.16	0.90	**1.55**	0.94	**1.09**

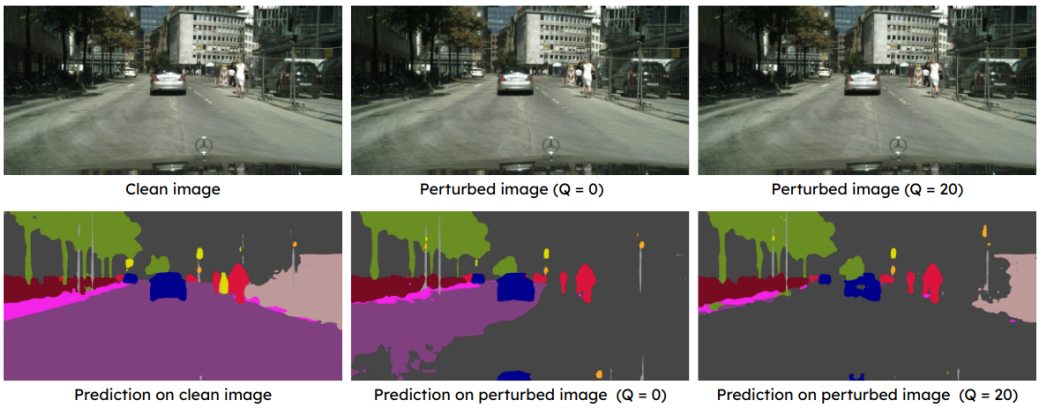


(a) Original Prediction **(b)** Least-likely Prediction **(c)** Attack Target

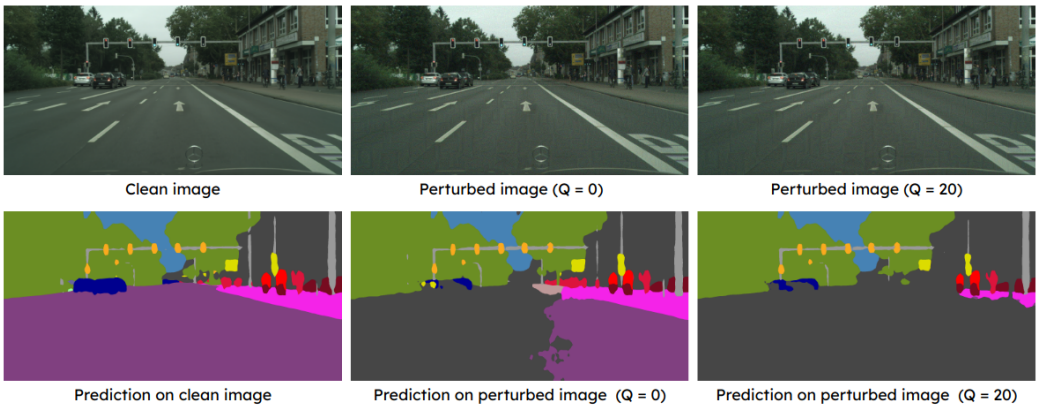
Figure 3.9: Our segmentation targeted attack setup. We select an object region y in the original prediction from surrogate `FCN` (Figure 3.9a). Identify the targeted label $y^\star$ from Figure 3.9b and craft the attack target Figure 3.9c. Image id: `frankfurt_000001_007857`

Table 3.12: Semantic segmentation targeted pixel success ratio (PSR) (%) for blackbox victim models and whitebox surrogate models with different ensemble sizes (N). The higher value indicates better attack performance. Experiment with CityScapes dataset, $\ell_\infty \leq 8$. PSP-r50, PSP-r101, DL3-r50, DL3-r101 stands for `PSPNet` and `DeepLabV3` built on `ResNet50`, `ResNet101` backbone respectively.

N	**Surrogate Ensemble**					**Blackbox Victim Models** (PSR ↑)			
	FCN	UPerNet	PSANet	GCNet	ANN	PSP-r50	PSP-r101	DL3-r50	DL3-r101
1	69.51	-	-	-	-	39.15	10.21	35.02	7.58
2	84.62	89.30	-	-	-	83.97	51.80	82.70	46.95
3	79.64	85.48	82.89	-	-	88.88	64.63	85.55	60.88
4	83.82	88.50	87.00	88.12	-	91.51	64.28	87.19	63.88
5	86.55	91.10	89.75	90.00	87.82	**92.91**	**69.09**	**88.95**	**69.65**



(a) Generate perturbation to map the Road region to Building. We use the $N = 2$ surrogate ensemble to generate this attack against blackbox victim `PSPNet-Res50`. For the direct transfer attack (at $Q = 0$), only 67.08% of the pixels of the target region are successfully mapped to the desired class. After weight optimization (with $Q = 20$), pixel success rate increases to 99.77%. Image id: `frankfurt_000001_007857`



(b) Generate perturbation to map the Road region to Building. We use the $N = 6$ surrogate ensemble to generate this attack against blackbox victim `DeepLabV3-Res50`. For the direct transfer attack (at $Q = 0$), only 63.06% of the pixels of the target region are successfully mapped to the desired class. After weight optimization (with $Q = 20$), pixel success rate increases to 99.98%. Image id: `munster_000003_000019`

Figure 3.10: (Caption next page.)

Figure 3.10: Visual adversarial examples of our method for targeted attacks to fool a blackbox semantic segmentation model.

3.4.8 Runtime and resource usage

We performed experiments on a single RTX 3090 GPU. Average time per query to attack an object detector for a 375×500 image with ensemble size $N = \{2, 5\}$ is $\{0.5, 1\}$sec. Average time per query to attack a segmentation model for a 512×1024 image with ensemble size $N = \{2, 5\}$ is $\{2.5, 5.5\}$sec.

3.5 Conclusion

We presented a new method to generate targeted attacks for dense prediction task (e.g., object detectors and semantic segmentation) using an ensemble of surrogate models. We demonstrate that (victim model-agnostic) weight balancing and (victim model-specific) weight optimization play a critical role in the success of attacks. We present an extensive set of experiments to demonstrate the performance of our method with different models and datasets. Finally, we show that our approach can create adversarial examples to fool multiple blackbox models and tasks jointly.

Limitations. Our method employs an ensemble of surrogate models to generate attacks, which inevitably incurs higher memory and computational overhead. Moreover, the success of our method hinges on the availability of a diverse set of surrogate models, which could potentially limit its efficacy if such models are not readily obtainable.

Chapter 4

Context-Aware Transfer Attacks for Object Detection

4.1 Introduction

Generating adversarial attacks (and defending against such attacks) has recently gained a lot of attention. An overwhelming majority of work in these areas have considered cases when images contain one predominant object (e.g., ImageNet [120] data), and the goal is to perturb an image to change its label. In real-life situations, we usually encounter images with many objects. Object detectors take a holistic view of the image and the detection of one object (or lack thereof) depends on other objects in the scene. This is why object detectors are inherently context-aware and adversarial attacks are more challenging than those targeting image classifiers [165, 347, 79, 309].

In this paper, we focus on the problem of generating context-aware adversarial attacks on images to affect the performance of object detectors. Our approach is to craft an **attack plan** for

each object, which not only perturbs a specific **victim object** to the target class, but also perturbs other objects in the image to specific labels or inserts phantom objects to enhance the holistic context consistency; these associated objects are called **helper objects**. The helpers are selected based on the **context graphs**, which capture the co-occurrence relationships and relative location and size of objects in the image. The context graphs can be learned empirically from natural image datasets. The nodes of a context graph are object classes, and each edge weight captures the co-occurrence, relative distance, and size likelihood of one object conditioned on the other. The intuition is that each class is often associated with certain classes, and unlikely to be associated with certain others.

Our interest lies in blackbox attacks where the perturbations generated for an image are effective on a variety of detectors that may not be known during the perturbation generation process. The conceptual idea of our proposed approach is to generate perturbations with an ensemble of detectors and subsequently test them on an unknown detector. Such attacks are referred to as transfer attacks, and we refer to the unknown detector we seek to fool as the **victim blackbox model**. To achieve this goal, we propose a novel sequential strategy to generate these attacks. We sequentially add perturbations to cause the modification of the labels of the victim and helper objects, based on the co-occurrence object relation graph of the victim object. This strategy is the first to use explicit context information of an image to generate a blackbox attack plan. Note that the sequential strategy makes a small number of queries (2–6 in our experiments) to the blackbox detector as new helper objects are added in the attack plan. The blackbox detector provides hard labels and locations of detected objects. We use this information only as a stopping criterion for the attack generation, unlike query-based approaches [485] that often need to use thousands of queries to estimate local gradients. The framework is illustrated in Figure 7.1.

The main contributions of this paper are as follows.

- This is the first work that considers co-occurrence between different object classes in complex images with multiple objects to generate transferable adversarial attacks on blackbox object detectors.
- We show how to generate context-aware attack plans for targeted mis-categorization attacks. The attacks generated using context-aware attack plans provide significantly better transfer success rates on blackbox detectors than those generated by methods that are agnostic to context information (an average improvement of more than 10 percentage points in fooling rate, see Table 4.1).
- Our comprehensive evaluations also include context-aware adversarial attacks on multiple datasets using multiple object detectors. We also provide analysis on the effect of helper objects in generating successful attacks and the generalizability of contexts.

4.2 Related Work

Context in object detection. The importance of context has been studied extensively to enhance visual recognition technologies [433, 454, 128, 155, 329, 522, 351]. Modern object detectors [396, 393, 76] consider holistic information in the image to locate and detect different objects, and several works explicitly utilize context information to improve the performance of object detectors [37, 535, 100, 311, 33, 474]. Some recent papers have considered context consistency to detect adversarial attacks [288, 525], but the attack generation uses existing whitebox attack schemes that do not consider context information explicitly. To the best of our knowledge, we are the first to use context information of objects explicitly for generating attacks on object detectors for images with multiple objects.

Blackbox adversarial attacks. Blackbox attacks is a practical setting where the attacker can only query the model and get the output instead of having access to the model's internal parameters. Two common strategies targeting this challenging problem are transfer-based attacks and query-based attacks. Query-based attacks have high success rates but require an overwhelmingly large number (often hundreds or thousands) of queries [56, 92, 177, 208, 101, 88, 278, 485]. In this paper, we explore a more stringent case where only a very small number of model calls is allowed. Several papers [373, 309, 131, 281] have examined the phenomenon of transfer attacks where the adversarial examples generated using a surrogate network can fool an unknown network. The previous works studying transfer attacks focus on image classifiers. In this paper, we focus on object detectors, which is considered to be a much harder problem [506, 500].

Attacking object detectors. Almost all existing attacks on object detectors focus on whitebox setting. Some patch-based attacks [307, 406] are very effective but the patches are obviously visible to observers. Some attacks such as DAG [506], RAP [294], and CAP [533] rely on region proposal network (RPN), thus only work for proposal-based (two-stage) object detectors. Some attacks are more generic such as UAE [490] and TOG [106] that work for both one-stage and two-stage models. Among them TOG is the most generic approach that can attack all different kinds of models regardless of their architectures as long as backpropagation on training loss is feasible. Even though some of these works have reported transfer attack results on a small set of blackbox models, since they are mainly designed for whitebox attacks, they fail to provide systematic evaluation in a realistic blackbox settings.

4.3 Context-Aware Sequential Attacks

While algorithms in prior approaches search for adversarial examples that misclassify the victim objects only, we propose to formulate the optimization problem towards perturbing both the victim object and the "context" associated with the victim object. The context of an object is determined by the objects that co-exist with it. We hypothesize that the context not only plays an important role in improving classification/detection performance, but can also boost the ability to realize efficient adversarial attacks against object detectors.

Next, we show how we compose context-aware attack plans and search for adversarial examples by sequentially solving optimization problems that are defined for a context-aware attack plan. The context-aware attack plans utilize the contextual information with regard to the co-existence of instances of different categories and their relative locations and sizes. We first describe how we represent the contextual information. Then we discuss how to compose the so-defined context-aware attack plan. Finally, we describe how we generate the adversarial examples by solving relevant optimization problems sequentially. The framework is explained in detail in Figure 7.1.

4.3.1 Context Modeling

We represent a natural scene image as I and the distribution of all natural images as $\mathcal{D}$. Each $I \in \mathcal{D}$ could contain one or multiple object instances. We denote the possible object categories in the distribution $\mathcal{D}$ by $C = \{c_1, c_2, ..., c_k\}$, where k is the total number of object categories. We define the context graph (an example shown in Figure 7.1 b) as a fully connected directed graph, in which each node is associated with an object category c_i and the weight on the edge $e_{i,j}$ encodes

different properties relating two nodes such as their co-occurrence probability, distance, and relative size. The number of nodes in the context graph is same as the number of object categories k.

Co-occurrence graph. We aim to model the co-occurrence probability of each pair of instances in a natural image. To be more specific, we seek to determine the probability of the event that an instance of category c_j appears in the image *given that* an instance of category c_i also appears in the image. Co-occurrence graph inherits the structure of the aforementioned context graph, and the directed edge $e_{i,j}$ represents the probability that an instance of category c_j appears in the image given an instance c_i already exists. This probability is denoted by $p_{i,j}^{\text{occur}} = p(c_j|c_i)$. Note that for each node, we also have an edge pointing to itself (i.e., $e_{i,i}$). According to the definition of probability we have $0 \leq p_{i,j}^{\text{occur}} \leq 1$ and $\sum_{j=1,\ldots,k} p_{i,j}^{\text{occur}} = 1$ for all i.

To compose such a co-occurrence graph, we can calculate a matrix $P = \{p_{i,j}^{\text{occur}} | i, j = 1, \ldots, k\}$ using a large-scale natural scene image dataset $\mathcal{D}'$, whose distribution is deemed to be similar to $\mathcal{D}$. We approximate co-occurrence probabilities using the relative co-occurrence frequencies of objects from $\mathcal{D}'$.

Distance graph. Suppose the bounding box of an object is given as $[x^c, y^c, h, w]$, where (x^c, y^c) denote the center pixel location of the box and (h, w) denote the height and width in pixels. In distance graph, the edge $e_{i,j}$ captures the distribution of the ℓ_2 distance between center points of c_j and c_i. The bilateral edges are equivalent. Considering the fact that the image size (H, W) varies in the dataset, which will also influence the distance between two objects, thus to minimize this scaling effect, we normalize the distance by image diagonal $L = \sqrt{H^2 + W^2}$. The distance distribution is denoted as $p_{i,j}^{\text{dist}}(\ell_2([x_i^c, y_i^c], [x_j^c, y_j^c])/L|c_i)$.

Size graph. Similarly, size graph models the 2D distributions of object height and width, where edge $e_{i,j}$ represents $p^{\text{size}}_{i,j}(h_j/L, w_j/L|c_i)$, which is the distribution of height and width of c_j given c_i is also present in the image.

4.3.2 Context-Aware Attack Plan

Given an image I, we denote the instance categories in the image as $X = [x_1, x_2, ..., x_m]$, where m is the total number of detected objects in the image. Note that different x_i could be the same because two instances of the same category can co-occur in a scene.

For the miscategorization attack, the goal is to miscategorize x_i to x'_i for $i \in \{1, \ldots, m\}$. We call the object associated with x_i as the **victim object** or the **victim instance**. To simplify our discussion, let us assume that our goal is to miscategorize x_1 to x'_1. Note here that methods that focus on miscategorizing the victim object/instance only will search for a perturbation so that the labels for all the objects become $X' = [x'_1, x_2, ..., x_m]$. We call X' the **attack plan**, since it yields the target labels for attackers.

In our proposed context-aware attack method, in addition to miscategorizing x_1 into x'_1, we may also want to miscategorize one or more helper objects that can provide important context information for x_1. We create a context-aware attack plan as $X'_c = [x'_1, x'_2, ..., x'_n]$. We will use subscript c with context-aware attack plans to distinguish them from the context-agnostic attack plans. The x'_i could be the same as x_i when we do not seek to miscategorize the instance associated with x_i. All the x_i (except the victim object) that change to a different label x'_i in X'_c are called **helper instances**. Note that in the context-aware attack plan, X'_c, n could be greater than m in cases where we decide to insert new instances as helper objects. We illustrate an example attack plan in Figure 7.1(c), where the bird at the bottom is the victim object that we want to mis-categorize to a

table; the bird at the top (to be mis-categorized as chair) and the additional appearing chair are the two helper instances.

The number of helper instances is a hyper-parameter that we tune. We use the co-occurrence graph, defined previously, to decide which existing instances (x_i) should serve as helper instances, and what category labels (x_i') should be assigned to these instances. From the co-occurrence graph, we obtain the co-occurrence probability with respect to every possible instance pair category. Given the goal of miscategorizing victim object from x_1 to x_1', we choose the label for every helper instance (x_i') by sampling the label space C according to the co-occurrence probability $p(x_i' = c|x_1')$ for all $c \in C$. Note that $\sum_{c \in C} p(x_i' = c|x_1') = 1$. We could model the joint probability of all helper instances given the target label, but that would require a large amount of data. Our sampling approach assumes conditional independence of helper instances (akin to naïve Bayes), in which we draw the most probable labels for our helper labels by sampling one row of the co-occurrence probability matrix. By random sampling the label space in this manner, we expect that objects that occur more frequently will be selected as labels for the helper objects. We first select the helper objects from among the m objects present in the scene. In case we need to add new helper instances $(> m)$, we choose their locations and sizes according to the mean values of the distributions given by distance and size graphs.

4.3.3 Sequential Attack Generation

We propose a sequential perturbation generation strategy, where we start with zero helper objects in the attack plan and sequentially add one helper object until the attack succeeds on the blackbox, as shown in Figure 7.1. We generate adversarial attacks using a single or multiple surrogate model(s) in our perturbation machine. As we sequentially add the helper objects in the attack

plan, we query the black-box model to see if our attack succeeds. In our experiments, we make up to 6 queries to the blackbox detector, which provides hard labels for the detected objects. We use this information only as a stopping criterion for the attack generation. We stop the sequential attack process if the adversarial example fools the black-box model or we run out of the budget of helper objects. Note that our strategy is orthogonal to query-based methods that aim to generate adversarial examples or estimate gradients of the blackbox models (often using hundreds or thousands of queries) [485, 101, 208].

Our attack generation method with a single surrogate detector is based on targeted adversarial objectness gradient attacks (TOG) [106], which can be viewed as training the detector for modified labels given in the attack plan X'. The weights of the detector network remain fixed but a perturbation image δ is added to the clean image as $I + \delta$ at every iteration to minimize the training loss $\mathcal{L}(\text{clip}(I + \delta); \mathcal{O}')$ for a desired output $\mathcal{O}'$. The value of $I + \delta$ is clipped at each iteration to make sure it is legally bounded. We generate the desired output $\mathcal{O}'$ based on our attack plan X'. The attack plan in X' only contains label information, but we also assign location and confidence score information in $\mathcal{O}'$. At every iteration, we update the perturbation using the iterative fast gradient signed method (I-FGSM),

$$\delta \leftarrow \delta - \epsilon \cdot \text{sign}[\nabla_\delta \mathcal{L}(\text{clip}(I + \delta); \mathcal{O}')], \tag{4.1}$$

where ϵ is the step size at each iteration. We can also use an ensemble of detectors as the surrogate models in perturbation machine, where we generate perturbation by minimizing the joint loss function over all detectors:

$$\mathcal{L} = \alpha_1 \mathcal{L}_1 + \alpha_2 \mathcal{L}_2 + ... + \alpha_N \mathcal{L}_\text{N}, \tag{4.2}$$

while keeping $\sum_{i=1}^{N} \alpha_i = 1$ and $\alpha_i > 0$ for all i. We can easily modify our method to use other perturbation generation methods and loss functions [326, 79, 131, 508, 300, 484].

4.4 Experiments

We perform comprehensive experiments on two large-scale object detection datasets to evaluate the proposed context-aware sequential attack strategy. We mainly show that the context-aware sequential attack strategy can help with mis-categorization attacks in blackbox setting. We also present results with whitebox setting, for completeness, even though this is not our primary objective.

4.4.1 Implementation Details

Object detection models. We evaluate our attack plans on a diverse set of object detectors, including

- **Two-stage detectors.** Faster R-CNN [396], Libra R-CNN [370, 369];
- **One-stage detectors.** YOLOv3 [393], RetinaNet [301];
- **Anchor-free detectors.** FoveaBox [257], FreeAnchor [539];
- **Transformer-based detectors.** DETR [76], Deformable DETR [552].

We use `MMDetection` [90] code repository for the aforementioned models. Inspired by [309, 500], we use an ensemble of locally trained object detection models as the surrogate model.

Perturbation Budget	Method	**Whitebox**		**Blackbox**					
		FRCNN	YOLOv3	Retina	Libra	Fovea	Free	DETR	D-DETR
				Results on PASCAL VOC					
$L_\infty \leq 10$	Baseline	40.0	53.8	13.8	9.2	22.2	27.4	9.6	23.2
	Random	52.4	69.2	19.4	17.4	31.6	37.8	17.4	36.8
	Ours	**55.8**	**75.6**	**22.6**	**20.4**	**33.6**	**39.2**	**20.2**	**39.2**
$L_\infty \leq 20$	Baseline	65.2	67.8	24.0	21.4	34.4	41.8	14.4	37.6
	Random	74.4	83.8	31.0	29.6	46.2	54.4	28.0	52.6
	Ours	**78.6**	**87.2**	**35.2**	**38.4**	**51.6**	**56.6**	**34.0**	**58.4**
$L_\infty \leq 30$	Baseline	70.6	70.4	29.8	28.6	41.6	48.0	20.4	38.6
	Random	79.2	82.6	37.8	36.8	53.4	59.8	34.4	52.8
	Ours	**80.6**	**88.0**	**42.0**	**44.2**	**56.8**	**63.6**	**40.2**	**59.0**
				Results on MS COCO					
$L_\infty \leq 10$	Baseline	29.0	32.2	7.4	4.8	11.6	16.6	3.4	19.0
	Random	40.2	48.4	11.2	8.0	14.6	20.0	6.2	23.6
	Ours	**41.2**	**54.4**	**12.0**	**11.2**	**18.6**	**25.0**	**10.8**	**27.8**
$L_\infty \leq 20$	Baseline	51.8	49.2	13.4	11.8	22.0	28.6	8.8	26.8
	Random	60.6	66.4	20.6	18.8	31.4	37.2	**20.2**	39.2
	Ours	**64.4**	**70.0**	**20.8**	**22.2**	**35.4**	**40.8**	20.0	**43.2**
$L_\infty \leq 30$	Baseline	57.6	54.4	18.2	15.4	25.6	34.8	8.0	28.8
	Random	65.8	73.6	23.8	21.8	34.8	**47.8**	18.4	42.0
	Ours	**68.6**	**75.4**	**27.2**	**27.2**	**39.2**	46.2	**21.2**	**48.6**

Table 4.1: (Caption next page.)

Table 4.1: White-box and black-box mis-categorization attack fooling rate on different models with different perturbation budgets ($L_\infty \leq \{10, 20, 30\}$) using VOC and COCO dataset. Baseline only perturbs the victim object, while ours also perturbs other objects conforming to context. Random perturbs other objects but assign random labels. **Abbreviation:** Faster R-CNN (FRCNN), RetinaNet (Retina), Libra R-CNN (Libra), FoveaBox (Fovea), FreeAnchor (Free), Deformable DETR (D-DETR).

Selecting a good surrogate ensemble is an interesting question, where the number and type of surrogate models will influence the attack success rate. We tested different single and multiple models as surrogates in our preliminary tests and observed a similar trend that the context-aware attacks significantly outperform the baseline attacks that are context-agnostic. Therefore, we selected two most commonly-used models, Faster R-CNN and YOLOv3, as the surrogate ensemble in our experiments. The weighting factor α is chosen such that the individual loss terms are balanced. On the blackbox victim side, we choose the leftover models that have a variety of different architectures.

Datasets. We use images from both PASCAL VOC [140] and MS COCO [302] datasets in our experiments. VOC contains 20 object categories which commonly appear in natural environment, and COCO contains 80 categories which is a super-set of the categories in VOC. We randomly selected 500 images that contain multiple $(2-6)$ objects from `voc2007test` and `coco2017val`. Since all models in `MMDetection` are trained on `coco2017train`, while testing the detectors on VOC images, we only return the objects that also exist in VOC categories.

Context graph construction. For VOC and COCO images, we extract context from `voc2007trainval` and `coco2017train` respectively. For each dataset, we build three $N \times N$ arrays (N is number of labels) that contain co-occurrence probability, distance distribution, and size distributions. The (i, j) cell in the co-occurrence array stores the number of co-occurrences of object c_i and object c_j normalized by the summation of that row; each cell in the distance array is a 1D distribution of the distances between c_i and c_j found in the images; each cell in the size table is a 2D distribution of h and w of c_j given c_i. These three arrays can be easily computed form the datasets.

Attack generation. We use I-FGSM-based method to generate a perturbation on the whole image (as discussed in Eqn. (4.1)), and we limit the maximum perturbation level to be $L_\infty \leq \{10, 20, 30\}$. The number of helper objects is empirically chosen to be 5. We present an analysis study on how the attack performance changes with the number of helper objects in Section 4.4.3 of analysis study.

Baseline and comparisons. TOG [106] shows better performance compared to UEA [490] and RAP [294]; therefore, to understand the performance of the proposed context-aware attack plan strategy, we use the current state-the-art attack strategy based on TOG [106]. The attack plan generated by the baseline (labeled as Baseline in Table 4.1) is context-agnostic and only associated with the victim object. To validate that our proposed context-aware attack really benefits from co-occurrence, location and size information, we also present results for a setting (labeled as Random in Table 4.1) in which we choose helper objects label and location at random.

Evaluation metric We use attack success rate (or fooling rate) to evaluate the adversarial attack performance on any victim object detector. Since we perform targeted mis-categorization attack,

instead of using mAP which takes account of all existing objects, we only focus on the victim object and define our attack success rate as the percentage of attacks in which the victim object was successfully mis-classified to the target label. In experiments, we check if the target object exists in the detection with an intersection over union (IOU) greater than 0.3. If yes, the attack is successful (or the detector is fooled); otherwise, the attack fails. For the selection of target objects, we randomly selected one target label that is not present in the original image to mimic the out-of-context attack as well as eliminating the chance of miscounting the existing objects as success.

4.4.2 Evaluation of Attack Performance

Whitebox attack performance. We observe that the attack success rate suffers even in whitebox setting, especially when the perturbation budget is small. As shown in Table 4.1, the baseline whitebox attack with $L_\infty \leq 10$ on COCO can only achieve around 30% fooling rate. This is because we simultaneously attack multiple objects in the image and also use an ensemble loss to fool multiple models jointly, the targeted mis-categorization attack is challenging. Even in this difficult setting, our context-aware attack can successfully improve the fooling rate by $10 - 20$ percentage points. Besides this, we can observe that our method provides significant improvement (by at least 10 percentage points) over the baseline method at all perturbation levels on both VOC and COCO dataset. Our performance is not only better than baseline method, but also has clear advantage over sequential attacks with random context. This validates the effectiveness of the proposed context-aware sequential attack strategy in the whitebox settings.

Blackbox attack performance. We test the performance of the attacks generated by the surrogate detectors in the perturbation machine on different blackbox detectors. Our hypothesis is that

the context-aware adversarial examples transfer better to the unseen models, and thus have better attack performance compared to context-agnostic (baseline) attacks in the blackbox setting. We use the same baseline and evaluation metrics as in the evaluation of the whitebox attack. Our results corroborate our hypothesis as we observe that even though the blackbox attack success rate is significantly lower compared to the whitebox attack success rate, our proposed context-aware sequential attack strategy still provides significantly better transfer success rate compared to the context-agnostic (baseline) attacks. For both VOC and COCO datasets, for all levels of perturbation. Overall, for every test setting, our method improves the success rate over baseline method by 5–20 percentage points (average improvement is beyond 10 percentage points). This is a significant improvement for the notoriously difficult problem of transfer attacks on object detectors in blackbox settings by using just 2–6 queries. Our proposed context-aware attack strategy has better transfer rates than the context-agnostic baseline and random assignment of labels, which further shows the benefits of utilizing co-occurrence relationships, location and size information to generate the attack plans.

Visualization. We show three attack examples in Figure 4.2. In the first example, we aim to miscategorize a TV monitor to sofa. We observe that the baseline attack fails to transfer to the blackbox model, RetinaNet (middle row). In comparison, the context-aware adversarial example from our method fools the victim blackbox model to detect the TV monitor as a sofa by introducing a pottedplant as the helper object, which frequently co-occurs with the target label, sofa. In the second example, we aim to miscategorize a person into a bird. The baseline attack fails since the person is still detected. However, our method succeeds by introducing another bird as the helper object. In the third example, we seek to mis-categorize a cow as a sofa. The baseline attack fails as no object

is detected near the victim object. Our context-aware attack plan succeeds by assigning the person and other cow in the image to chairs (helper objects).

4.4.3 Analysis Study

Number of helper objects. Even though helper objects boost the adversarial attack success, we do not need a large number of them. Since the perturbation budget is fixed, using too many helper instances may reduce the effect for the victim instance. On the other hand, not using any helper instances would completely eliminate the benefits of context-aware attacks. To investigate how the number of helper objects affects the attack performance, we plot the mis-categorization attack success rate with respect to the number of helper objects in Figure 4.3. We observe that adding more objects improves attack success rate both for the whitebox and blackbox attacks. The improvement is profound for some blackbox attacks that have low baseline attack success rates. We also observe that the first few helper objects boost the attack performance significantly and the improvement gradually plateaus as we add 4–5 helper objects.

Context graphs of different datasets. To demonstrate that the context graphs are generic enough to be used across different natural scene datasets, we evaluate the similarity of the co-occurrence matrices extracted from the two large-scale datasets (VOC and COCO). The average Perason correlation coefficient of each corresponding row of VOC matrix and COCO matrix is 0.90, which signifies strong positive correlation between co-occurrence relationships encoded by these two context graphs. We can visually see the similarities of these two co-occurrence matrices in Figure 4.4. One of the salient patterns common in these two matrices is that the column of `person` is colored in dark green, showing that person generally has a high probability to co-occur with other objects.

This is a notable feature of natural scene images. Because of the high similarity of the contexts in the two datasets, we can use their context graphs interchangeably. It is indeed possible that if the original context of objects in the given image is very different from the context graph we use to build the attack plan, the transfer attack success rate will suffer. This can be corroborated by the comparison of Random and Ours in Table 4.1.

4.4.4 Analysis on Number of Helper Objects

We present additional results for perturbation levels $L_\infty \leq 10, 30$. We observe a similar trend as in Figure 4.3 that success of mis-categorization attacks increases as we add helper objects in our attack plans. In some cases, the success rate almost doubles compared to baseline as we add 5 helper objects.

4.4.5 Visualization Examples

We present some additional images to show comparison between our context-aware attack method with baseline method. We show examples where the perturbations generated by our method can successfully transfer to the blackbox model while the perturbations generated by baseline method fail. The experiment settings are the same as Figure 2 in the main paper.

4.5 Conclusion

In this paper, we propose a novel context-aware adversarial attack method that exploits rich object co-occurrence relationships plus location and size information to effectively improve mis-categorization attack fooling rate against blackbox object detectors. Our experimental results on two large-scale datasets show that our attack success rate is significantly higher than baseline and comparing methods, which validates the effectiveness of our methods. The contextual relationships modeled by our method holds true in different datasets within natural image domain, thus implying the wide applicability of our methods.

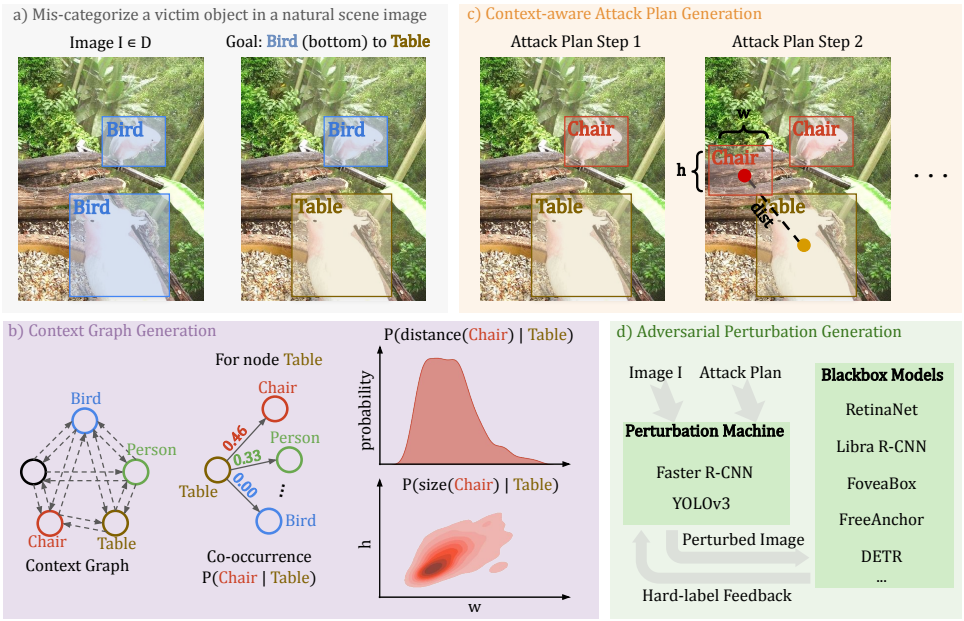


Figure 4.1: Overview of our framework for generating the context-aware sequential attack. a) Given a natural image, our goal is to trick an object detector to assign the victim object a given target label (e.g., bird to table). b) We construct a context graph that encodes the co-occurrence probability, distance, and relative size distribution relating pairs of objects (e.g., the edge from table to chair represents they co-occur with probability 0.46). c) Given the attack goal and context graph, we generate a context-aware attack plan that has a small number of steps. In each step, we assign target labels for existing objects and introduce new helper objects if needed. For example, co-occurrence of chair with table is most probable, we change the bird to a chair for stronger context consistency (depicted in Attack Plan Step 1). We may need to add a phantom chair around the table (as depicted in Attack Plan Step 2). d) Given the attack plan and the victim image, we generate perturbations using I-FGSM on the surrogate whitebox models in our perturbation machine. We test the perturbed image with the given blackbox model and based on the hard-label feedback, we either stop (when the attack is successful or when we exhaust our budget of the helper objects) or craft new attack based on the next steps and repeat the process.

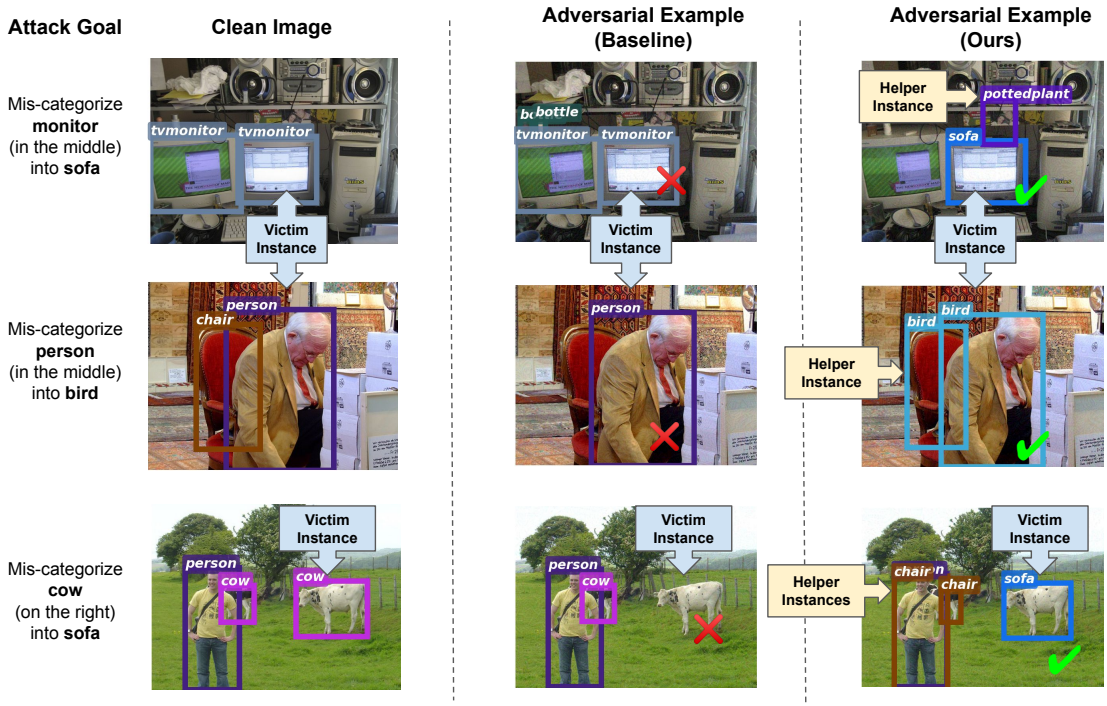
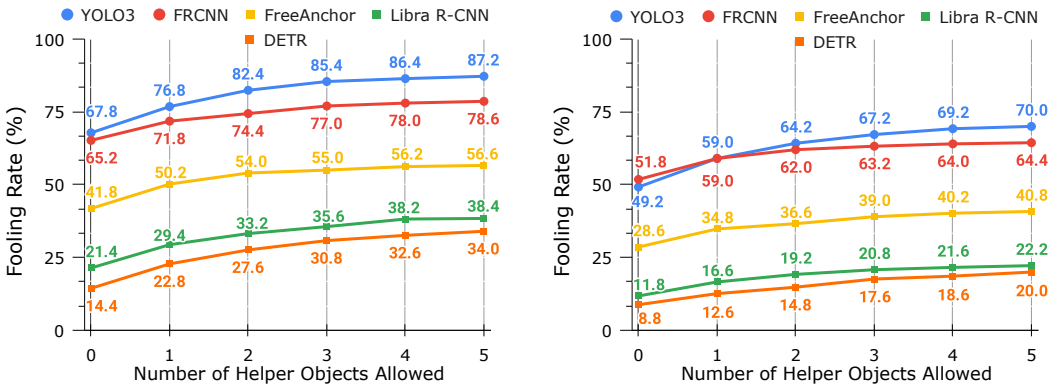


Figure 4.2: Examples where baseline attack fails but context-aware method succeeds by introducing helper objects in the attack. The perturbation ($L_\infty \leq 10$) is generated from our perturbation machine (whitebox ensemble of FRCNN and YOLOv3) and tested on the blackbox model (RetinaNet). The detection results on original image, image perturbed by baseline attack, and image perturbed by our context-aware method are shown in the subfigures from left to right. In these examples, we introduce `pottedplant` as a helper object to mis-categorize the victim `monitor` to `sofa`, introduce another `bird` to mis-categorize the `person` to a `bird`, and add a few `chairs` to mis-categorize the `cow` to a `sofa`. Visualization of perturbation level $L_\infty <= 20, 30$ can be found in supplementary materials.



(a) PASCAL VOC, $L_\infty \leq 20$ **(b)** MS COCO, $L_\infty \leq 20$

Figure 4.3: Mis-categorization attack fooling rate of white-box and black-box models at perturbation level $L_\infty \leq 20$ w.r.t. number of helper objects allowed (changed or added). Circles denote white-box models (FRCNN and YOLO3) and squares denote black-box models (FreeAnchor, Libra R-CNN, and DETR). Plots of perturbation level $L_\infty \leq 10, 30$ can be found in supplementary material.

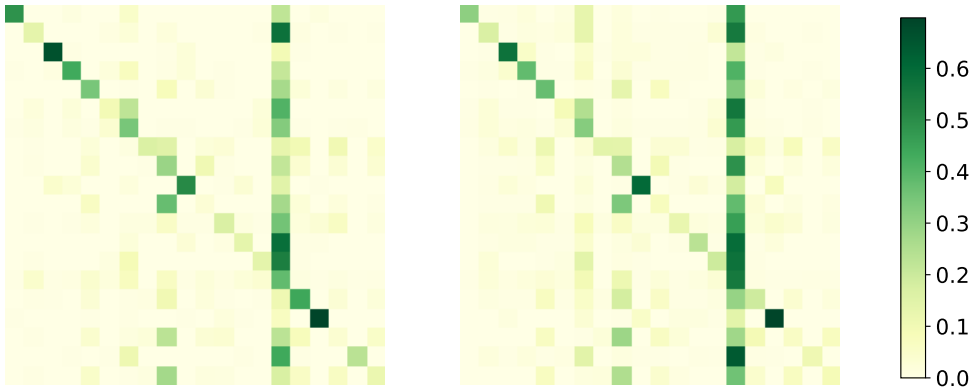
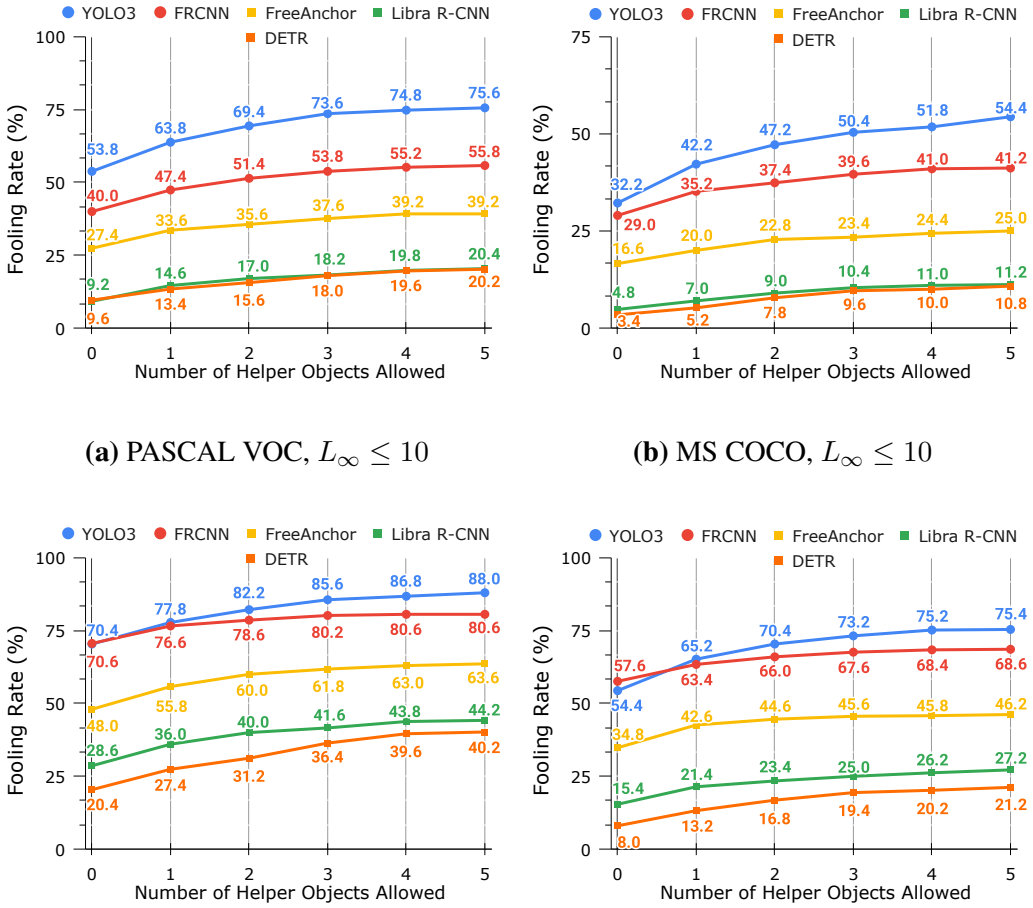


Figure 4.4: Co-occurrence matrices for VOC (left) and COCO (right) for 20 object categories that are common in both datasets.



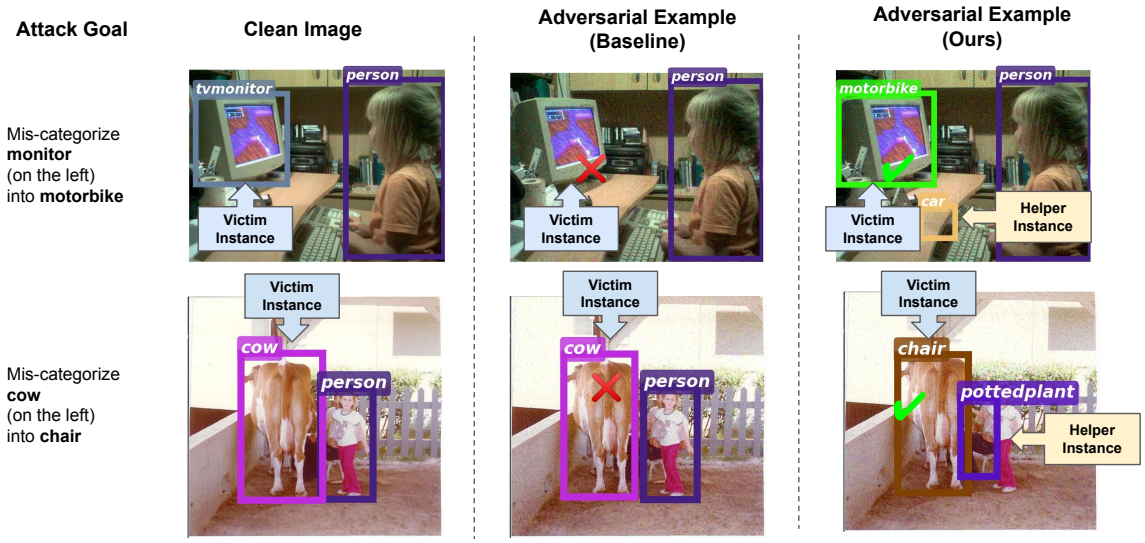
(a) PASCAL VOC, $L_\infty \leq 10$

(b) MS COCO, $L_\infty \leq 10$

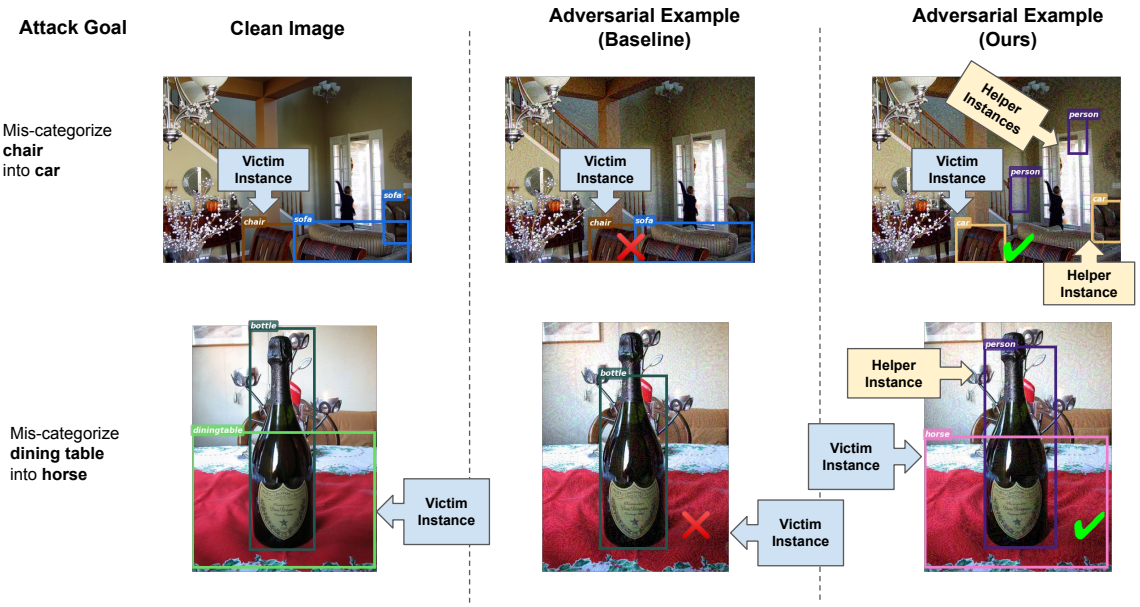
(c) PASCAL VOC, $L_\infty \leq 30$

(d) MS COCO, $L_\infty \leq 30$

Figure 4.5: Mis-categorization attack fooling rate of white-box and black-box models at perturbation level $L_\infty \leq 10, 30$ w.r.t. number of helper objects allowed (changed or added). In the legend, circle denotes white-box models (FRCNN and YOLO3) and square denotes black-box models (FreeAnchor, Libra R-CNN, and DETR). Baseline is where no helper objects is allowed.



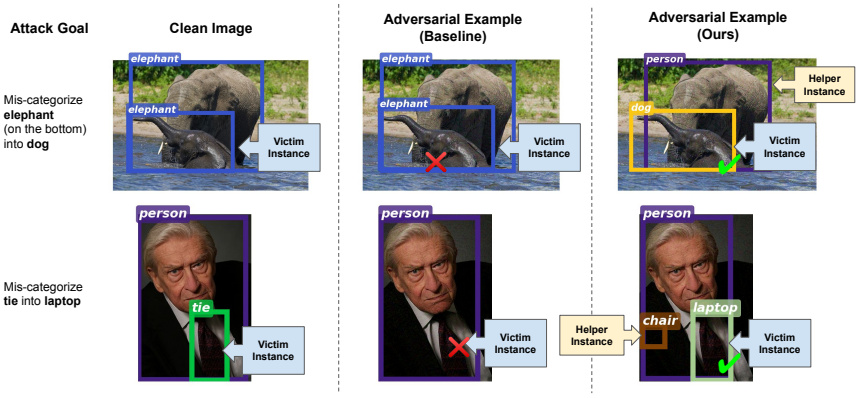
(a) PASCAL VOC, $L_\infty \leq 20$



(b) PASCAL VOC, $L_\infty \leq 30$

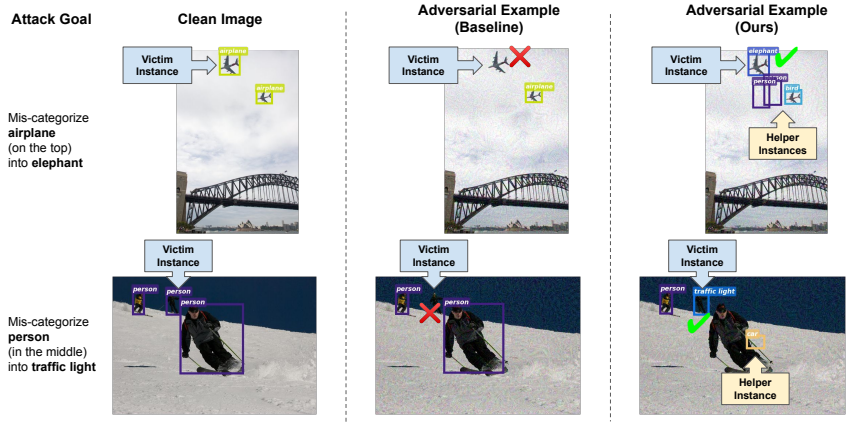
Figure 4.6: (Caption next page.)

Figure 4.6: Supplement to Figure 2, here we visualize four more examples under different perturbation budgets ($L_\infty \leq 20, 30$) where baseline attack fails but our context-aware method succeeds by introducing helper objects in the attack. The perturbation is generated from our perturbation machine (whitebox ensemble of FRCNN and YOLOv3) and tested on the blackbox model (RetinaNet). The detection results on original image, image perturbed by baseline attack, and image perturbed by our context-aware method are shown in the subfigures from left to right. In these examples, we introduce car as a helper object to mis-categorize the victim monitor to motorbike, introduce a potted plant to mis-categorize the cow to a chair, add a few persons and a car to mis-categorize the chair to a car, and change the bottle to a person in order to mis-categorize the dining table to a horse.



(a) MS COCO, $L_\infty \leq 10$

(b) MS COCO, $L_\infty \leq 20$



(c) MS COCO, $L_\infty \leq 30$

Figure 4.7: (Caption next page.)

Figure 4.7: Correspond to the previous visualizations on VOC dataset, here we also visualize examples for COCO dataset, where baseline attack fails but our context-aware method succeeds by introducing helper objects in the attack. The perturbation ($L_\infty \leq 10, 20, 30$) is generated from our perturbation machine (whitebox ensemble of FRCNN and YOLOv3) and tested on the blackbox model (RetinaNet). The detection results on original image, image perturbed by baseline attack, and image perturbed by our context-aware method are shown in the subfigures from left to right. In (a), we introduce a person as a helper object to mis-categorize the victim elephant to a dog, introduce a chair to mis-categorize the tie to a laptop; in (b), we add a few cows in the scene to mis-categorize the car to a cow, added an other donut to mis-categorize the stop sign to a donut; in (c), we perturb the airplane a bird and add a few persons to mis-categorize the airplane to an elephant, introduce a car to mis-categorize the person to a traffic light.

Chapter 5

Zero-Query Transfer Attacks on Context-Aware Object Detectors

5.1 Introduction

Despite achieving significant performance gains on a variety of vision and language tasks, deep neural networks (DNNs) are vulnerable to adversarial attacks [438]. One of the most popular adversarial approaches is the class of perturbation-bounded evasion attacks [165, 373, 79, 326]. Here, an attacker can make a model yield arbitrarily wrong classification results by adding imperceptible perturbations to the input image. These attacks are quite practical and can be performed at test time without needing access to the training data. The vast majority of work in this area has focused on attacking classifiers trained on datasets like ImageNet, MNIST, CIFAR-10, and CIFAR-100, where the classifier attempts to recognize one dominant object in a given image. In contrast, we are primarily concerned with object detectors [396, 391, 301, 257, 370] that localize and recognize

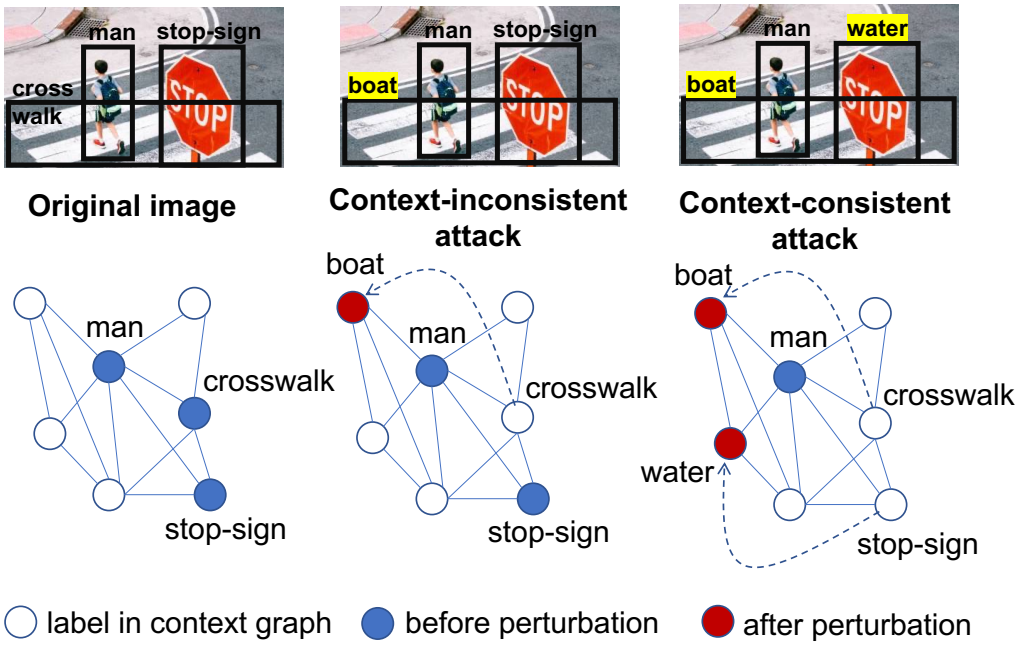


Figure 5.1: For natural scenes containing multiple objects, applying an evasion attack on an individual object (e.g., crosswalk → boat) violates the context: A boat and a stop-sign rarely occur together. A context-aware detector can detect this attack. In this work, we perturb multiple objects in a context-consistent way (e.g., crosswalk → boat, stop-sign → water) in a single attempt. The combination (person, boat, water) does not violate context and thus fools even a context-aware detector.

multiple objects in an image, which is the case in most natural images. Such detectors often take a holistic view of an image, rather than considering it as a collection of arbitrary objects [506, 500]. The objects in natural scene images form a context that can help identify the scene or given the scene we are likely to find the objects that conform to the scene context. For instance, a boat is unlikely to co-occur with a stop sign, and much more likely to co-occur with water. Leveraging this observation, some recent attack [66] and defense [288, 525] mechanisms have been proposed that

take image context into account. Context-aware defense methods in [288, 525] can detect attacks that are inconsistent with the scene context, as shown in fig:consistency-check. To evade these defenses, changing one object in the image is insufficient. The context-aware attack method in [66] uses the knowledge of co-occurrence between different object classes in complex images to generate a sequence of transferable attacks for black-box object detectors; however, this method needs a few queries to test which attack plan is successful. *We present, for the first time, a zero-query attack algorithm that changes multiple objects simultaneously in a context-consistent manner thereby creating a holistic adversarial scene that can overcome context-aware defenses.*

In this work, we consider "zero-query" attacks (ZQA) that refer to a setting in which the attacker has no feedback channel to access the classification decisions of the victim system. This setting is extremely useful in practice because in many applications the victim system is inaccessible to the attacker; even if the victim system is accessible, the attacker's communications can be monitored, and thus draw suspicion. ZQA, on the other hand, is a truly stealthy attack. The attacker can only implement an attack plan once by perturbing multiple objects in a given scene and submitting the perturbed image to the victim system. Furthermore, we assume that the victim system is explicitly context-aware; that is, it will examine the list of detected objects and determine whether that list is "context-consistent" or not. If yes, then the detector will not suspect an attack. If not, it will suspect that the image has been perturbed by an attacker. Our ZQA approach is able to subvert more sophisticated multi-label object detectors that either implicitly or explicitly take context relationships across objects into account while performing their inference. In fact, accounting for context is what makes it possible to achieve high success rates in a single attempt.

Several approaches exist for scene context modeling [194, 367, 453, 137, 33]. In this paper, we restrict our attention to object co-occurrence, which is the most fundamental approach to modeling semantic context. The context model is represented by a co-occurrence graph (or equivalently the co-occurrence matrix) that is computed for a given set of images. We consider a list of objects as context-consistent only if the corresponding labels form a fully connected sub-graph within the co-occurrence graph.

The main contributions of this paper are as follows.

- We develop an architecture for designing attacks on multi-object scenes that fool context-aware object detectors. Our detectors explicitly use object co-occurrence to model scene context.
- We propose an approach for zero-query context-aware attacks that generate adversarial scenes to fool a context-aware detector in a single step.
- We introduce the concept of a perturbation success probability matrix (PSPM) that models the probability of successfully perturbing a given target object to a given victim object in the white box setting. We use the PSPM to refine our attack plans, essentially choosing the one which is most likely to succeed. We show that the PSPM-guided attacks improve the fooling rate even in a black-box setting.
- We show experimentally that the fooling rate of ZQA is significantly higher than that achieved by a context-agnostic black-box attack. Furthermore, we compare our results against a possible "few-query" strategy [66] that repeatedly enhances the attack plan, while observing the detector output, until the detector is fooled. For the Pascal VOC dataset [140], the ZQA attacks provide fooling rates comparable to 5-query and 3-query attacks in the white-box and black-box settings, respectively.

5.2 Background and Preliminaries

5.2.1 Context in Object Detection

The role of context in improving visual recognition tasks has been well studied [350, 408, 522]. The state-of-the-art object detectors locate and detect several objects in the scene based on holistic information in the image [396, 391]. Many of these methods explicitly utilize context information to improve the performance of object detectors [38, 100, 312, 34, 36, 473]. Co-occurrence-based contextual information derived from image pixels [288], object labels [66], and language models [525] have also been used to build a context-aware object detector. Different from the above contributions, to the best of our knowledge, this is the first paper that proposes a context-aware attack to fool a context-aware detector with zero queries.

5.2.2 Black-Box Attacks

In a black-box attack, the attacker has no access to the internal parameters of the model; thus, instead of generating the perturbed image by backpropagation, the attacker can only test a perturbed image on the victim model. In some cases, the attacker can observe the output but in many cases even that is not possible [283]. This renders query-based attacks inapplicable, which usually take an overwhelmingly large number (often hundreds or thousands) of queries [56, 92, 177, 88, 278, 485]. In this paper, we explore the most stringent case where no model queries are allowed. Such attacks will be extremely hard to detect, free from suspicion of repeated queries, and thus will be a more viable option for subversion. Several papers [373, 309, 131, 281] have examined the phenomenon of transfer attacks where the adversarial examples generated using a surrogate network can fool a black-box victim network. A large body of work exists on designing (perturbation-bounded)

evasion attacks for images containing one predominant object and evaluating how well they transfer in a black-box setting [165, 78, 372, 81, 426]. In this work, our goal is to fool object detectors for general scenes. This is considered a harder problem because of the need to perturb multiple objects [506, 500]. This difficulty is exacerbated by the need to preserve contextual consistency during the attack [288]. A "few-query" strategy for context-aware attacks was proposed in [66] that repeatedly enhances the attack plan, while observing the detector output, until the detector is fooled. Our goal in this paper is to develop context-aware attacks with zero queries.

5.2.3 Attacks against Object Detectors

Attacking object detectors is harder than attacking classifiers, since the attack must obfuscate the category as well as the location of one or more objects [506, 500]. Object detectors [396, 391] implicitly use contextual information – for instance, relationships between object pixels and background pixels – to increase the speed of inference. Researchers have exploited this fact by developing various kinds of adversarial patches [307, 407, 199] which do not overlap with the victim objects, but can still fool the detectors. Some other attacks [506, 533, 490, 106], perturbing the image globally, are also successful in fooling object detectors in a white-box setting. A recently proposed method has demonstrated the ability to transfer attack black-box object detectors by exploiting context information, but the victim models do not explicitly check for context-consistency and also the attack needs multiple queries [66].

5.2.4 Defense methods

Some representative defense mechanisms [395] for mitigating adversarial attacks include enhancing adversarial robustness of the model intrinsically through adversarial training [326, 455,

25] or strengthening model architectures [308, 507]; and destroy adversary externally through input transformation [178, 505] or denoising [514, 409, 334]. These defenses are context-agnostic. Some recent papers consider context-aware object detectors operating on natural multi-object scenes [288, 525]. Although these works use different notions of context from our work, they confirm that attacking a context-aware detector is more difficult.

5.3 Context-Aware Zero-Query Attacks

We describe ZQA for natural scenes. A high-level diagram of zero-query context-aware black-box attacks is shown in fig:overview. We first present a high-level description and later elaborate on the building blocks. The attacker determines a list of objects detected in the scene and uses the co-occurrence-based context model to derive several context-aware attack plans that perturb one or more target objects to their respective victim labels. Then, given the perturbation budget, the attacker refines the list using a pre-computed PSPM, which we will discuss in detail later. The result is an attack plan that consists of a list of (victim label, target label) pairs that are most likely to succeed in fooling the victim object detector. The attacker then uses an evasion attack algorithm to generate the perturbed scene according to the refined attack plan. The perturbed image is sent to a black-box classification / detection machine equipped with an explicit context-consistency detection mechanism. The attack is considered successful only if the victim object is successfully perturbed to the target label *and* the victim system's object detector does not find any context inconsistency in its list of detected objects. We now describe the building blocks of the attack in detail.

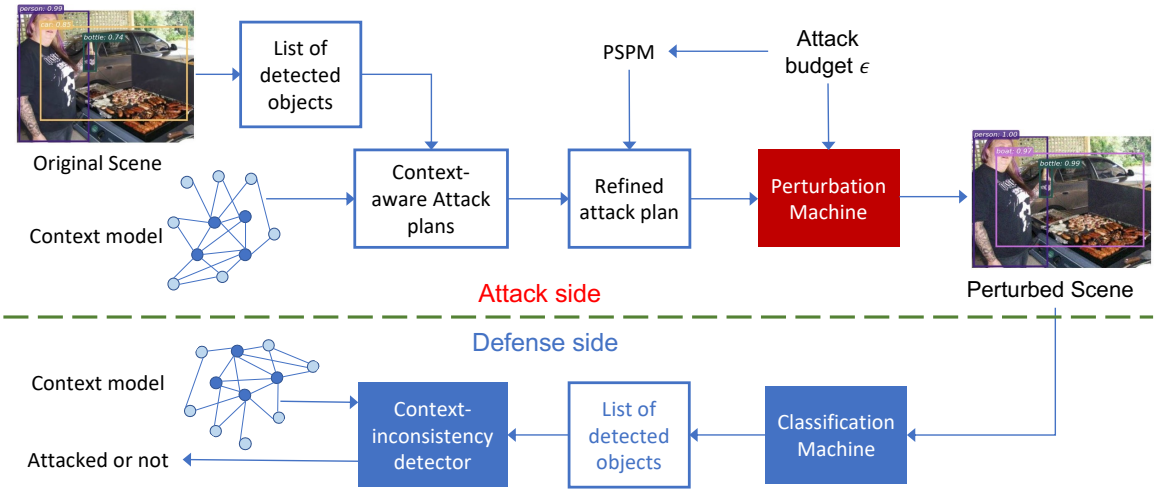


Figure 5.2: High-level diagram of zero-query context-aware black-box attacks. Given a victim image to be attacked, the attacker first finds a list of detected objects in the image and then consults the semantic context associated with the detector to design a context-aware attack plan that perturbs a victim object to a target label. To improve attack success rate, the attacker checks the PSPM corresponding to the perturbation machine with a certain perturbation budget ϵ and refines the original attack plan. With the refined attack plan, the attacker perturbs the image within bound ϵ, using the perturbation machine. The attacker's action is now complete. The perturbed image is then sent to a black-box classification / detection machine equipped with an explicit context-consistency detection mechanism. The attack is considered successful only if the victim object is successfully perturbed to the target object and the context-inconsistency detector does not find any inconsistency in the list of all detected objects.

5.3.1 Context Model

The context model is used by the attacker and the victim system's object detector to determine whether a given list of objects is context-consistent or not. Let us first define what is considered as context-consistent and what is context-inconsistent. Intuitively, combinations of objects detected in natural images from the training data should be considered as context-consistent, since these objects appear together in such scenes. On the other hand, combinations of objects should be considered as context-inconsistent if some of the objects have never appeared together in the training data. Thus, object co-occurrence is a fundamental cue in determining context-consistency.

Co-occurrence matrix/graph. We build a matrix, called the co-occurrence matrix $\mathbf{G}$, to model the co-occurrence relationships between objects as follows. Given a training data set with N labels and a label set $\mathcal{N} = \{\ell_1, \ldots, \ell_N\}$, a co-occurrence matrix $\mathbf{G}$ is an $N \times N$ matrix whose entries $\mathbf{G}(i, j)$ represent the number of unique pairs of objects with labels $\ell_i, \ell_j \in \mathcal{N}$ appearing together in the images. This matrix is symmetric before normalization. After normalizing each entry in $\mathbf{G}$ with the sum of elements in its row, we obtain $\mathbf{G}$, where $\mathbf{G}(i, j)$ indicates the conditional probability $p_{j|i}$ which is the probability of observing label ℓ_j if label ℓ_i is observed. In general, $\mathbf{G}$ is not symmetric. An example of $\mathbf{G}$ for the Pascal VOC Dataset is illustrated in fig:co-occurrence-matrix. The co-occurrence matrix can also be interpreted as a context graph (which we will also denote by $\mathbf{G}$) with N nodes, where the weight of the edge between nodes i and j represents the number of times that those two labels appear together in the training data.

Context consistency. If two nodes in a context graph do not have an edge connecting them, then these two labels never appear together in an image. Using this notion of co-occurrence-based con-

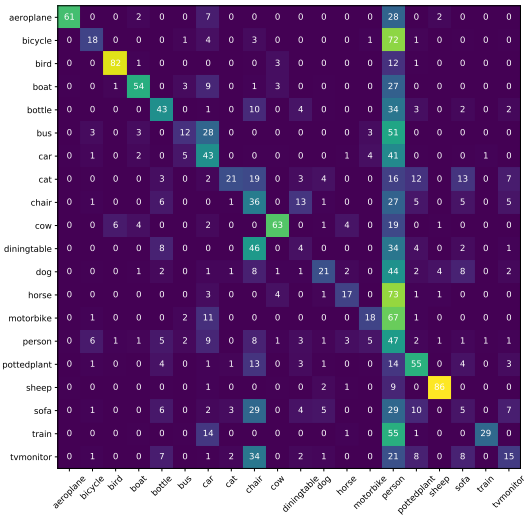


Figure 5.3: Co-occurrence matrix (conditional probability form) for the Pascal VOC07 training data set. Each cell indicates the probability as a quantized integer percentage (%).

text, we can define context-consistency and context-inconsistency as follows. A list of objects is considered as *context-consistent* if all the nodes representing the labels of those objects form a fully connected subgraph of the context graph $\mathbf{G}$. All the natural images in the data set satisfy this condition. A list of objects is considered as *context-inconsistent* if there are at least two nodes, representing the labels of two objects, that do not have an edge between them in the context graph.

Generalized context-consistency. Suppose we threshold the entries of $\mathbf{G}$ to obtain a matrix $\mathbf{H}_\eta$, where $\mathbf{H}_\eta(i,j) = \mathbf{G}(i,j)$ if $\mathbf{G}(i,j) > \eta$, for some threshold η. Otherwise $\mathbf{H}_\eta(i,j) = 0$. Build a context graph using $\mathbf{H}_\eta$, which we will also denote by the same symbol $\mathbf{H}_\eta$. Then the co-occurrence based notion of context consistency can be readily generalized as follows. A list of objects is considered as context-consistent up to a threshold η if all the nodes representing the labels of those objects form a fully connected subgraph of the context graph $\mathbf{H}_\eta$. Generalized context-inconsistency is defined similarly.

We describe the notion of context consistency for a list of object labels in fig:consistency-check. Let us take a clean image containing three objects as an example. Suppose we have a person, a crosswalk and a stop-sign in the image as indicated by the dark blue nodes. These three nodes form a fully connected subgraph indicating that these three objects are context consistent. Suppose we want to perturb the crosswalk in the image to a boat. An example of a context-inconsistent attack involves just perturbing the victim object (crosswalk) to the target label (boat denoted as a red node). The perturbed object list (person, boat, stop-sign) is no longer context-consistent because these three nodes of boat, person and stop-sign do not form a fully connected graph, indicating that the combination never appears in the training data. Intuitively, it would be unlikely to see a boat with a stop-sign in the natural images. An example of a context-consistent attack is to perturb the crosswalk to a boat but also perturb the stop-sign to water. That combination (person, water, boat) does appear in the training data, as seen from the fact that the label nodes form a fully connected sub-graph.

5.3.2 Perturbation Success Probability Matrix

Perturbation Success Probability Matrix (PSPM) is an $N \times N$ matrix denoted as $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}$ that is defined for an ensemble of classification models $\mathcal{C}$, perturbation budget ϵ, and an object perturbation algorithm α. PSPM is defined for a specific training data set with labels $\mathcal{N} = \{\ell_1, \ldots, \ell_N\}$. $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}(i,j)$ encodes the probability that an object with label ℓ_i can be successfully perturbed to label ℓ_j by the perturbation algorithm α using the perturbation budget ϵ.

The PSPM expresses an attacker's ability to perturb *individual* objects in a scene. The utility of the PSPM can be explained as follows. Suppose the list of objects in the given scene is $\mathcal{A} = \{A_1, \ldots, A_S\}$. Suppose the list $\mathcal{A}$ is perturbed to a list $\mathcal{B}$ using an evasion attack algorithm

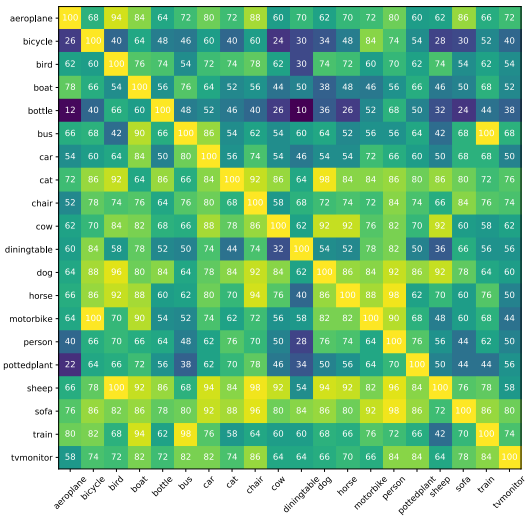


Figure 5.4: PSPM of Pascal VOC07 data set for $\mathcal{C} = \{\text{“Faster R-CNN”}\}$, $\epsilon = 10$, $\alpha = \{\text{“PGD}''\}$. Each cell indicates the probability as a quantized integer percentage (%).

α with perturbation budget ϵ. A context-agnostic attack would choose the labels in $\mathcal{B}$ at random from the label set $\mathcal{N}$. A context-aware attack would make sure that the labels in $\mathcal{B}$ are context-consistent as determined by the co-occurrence matrix $\mathbf{G}$. Then, there are in general one or more possible perturbation assignments $\mathcal{A} \rightarrow \mathcal{B}$ that are context-consistent. Depending upon the training set (e.g., the presence of objects in different poses, sizes, illuminations) some object perturbations, $A_i \rightarrow B_j$, are likely to be more successful than others, even in a white-box setting. Thus, each perturbation assignment $\mathcal{A} \rightarrow \mathcal{B}$ suggested by an examination of the co-occurrence matrix has a different likelihood of success. The PSPM enables us to select the assignment that is most likely to succeed. An example of a PSPM for the Pascal VOC07 dataset is shown in fig:pspm.

Some ways to choose an assignment amongst many assignments permitted by the co-occurrence matrix include

1. Choose the assignment $\mathcal{A} \rightarrow \mathcal{B}$ that maximizes each $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}(i,j)$ with $i \in \mathcal{A}, j \in \mathcal{B}$.

2. Choose the assignment $\mathcal{A} \rightarrow \mathcal{B}$ that maximizes the average of all $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}(i,j)$ with $i \in \mathcal{A}, j \in \mathcal{B}$.

3. Choose the assignment $\mathcal{A} \rightarrow \mathcal{B}$ that maximizes the minimum of all $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}(i,j)$ with $i \in \mathcal{A}, j \in \mathcal{B}$.

For simplicity, we use the first approach in the ensuing development. Admittedly, the PSPM considers the success of perturbing some $A_i \in \mathcal{A}$ to $B_j \in \mathcal{B}$ in a white-box setting; that is, for one or more classification models specified in the given ensemble $\mathcal{C}$. In a black-box setting, the attacker does not know which model is being used at the victim side. Thus, attack plans based on the PSPM are, at best, approximations. We hypothesize that these approximations are still better than choosing, at random, one of many possible assignments $\mathcal{A} \rightarrow \mathcal{B}$ suggested by the co-occurrence matrix. Our experiments, described in Section 5.4, indicate that refining the attack plan using the PSPM indeed improves the fooling rate compared to choosing an attack plan at random from the context-consistent candidate attacks suggested by the co-occurrence matrix.

5.3.3 Context-Consistent Attack Plan Generation

We now describe the zero-query method for deriving context-consistent attack plans using the co-occurrence matrix $\mathbf{G}$ and the PSPM matrix $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}$. The overall procedure is described in alg:oneshot. We assume that there is a desired target assignment for one object in the scene (e.g., change one of the horses to a bicycle). As discussed, just perturbing one object can result in a context-inconsistent list of objects in the perturbed scene. Thus, other objects in the scene may need to be perturbed so that the resulting list is context-consistent. The procedure described below ensures not only the attack plan is context-consistent, but it is also the plan that is most likely to

succeed in a white-box setting. As we mentioned earlier, for a black-box setting this plan may not always be the most likely to succeed. Experimentally, we found that the attack plan guided by PSPM also increases the fooling rate for black-box models.

5.3.4 Implementation of Attack Plan

To generate the adversarial scene, evasion attacks can be implemented using a single or multiple surrogate model(s). Our attack generation method with a single surrogate detector is based on projected gradient descent (PGD) [326] within a ℓ_∞ ball, which can be considered as a powerful multi-step variant of FGSM [165]. We initialize a zero perturbation $\delta^0 = 0$, and update it in each iteration as

$$\delta^{t+1} = \Pi_{\mathcal{S}} \left(\delta^t - \lambda \, \text{sgn}(\nabla_\delta \mathcal{L}(x + \delta^t, y)) \right), \tag{5.1}$$

where $\mathcal{L}$ is the loss function, $x \in \mathbb{R}^d$ is the input image and y is the target label. We generate the desired output y based on our attack plan, which includes object categories, locations, and confidence scores. We take a step size λ at each iteration t, and project (clip) $\Pi_{\mathcal{S}}$ the perturbation δ^t to the feasible set $\mathcal{S}$ which satisfies the following two criteria

$$\begin{cases} \|\delta^t\|_\infty \leq \epsilon, \\ x + \delta^t \in [0, 255]^d. \end{cases} \tag{5.2}$$

We use PGD for its simplicity in our experiments, but we can easily modify our method to use other (more advanced) perturbation methods such as MIM [131] and DIM [508] etc. without losing generalizability.

5.4 Experiments

We performed extensive experiments on two large-scale object detection datasets to evaluate the proposed context-aware zero-query attack strategy. We construct a query-based baseline scheme and evaluate the fooling rate of our proposed zero-query approach against it.

Datasets: We conduct our experiments using images from both PASCAL VOC [140] and MS COCO [302] datasets. VOC contains 20 common classes of objects, and COCO contains 80 classes which is a super-set of the categories in VOC. We randomly selected 500 images from `voc2007test` and `coco2017val` respectively, which contain multiple $(2-6)$ objects. This manuscript contains results for various models on the `voc2007test`. The results for the `coco2017val` are in the supplementary.

Attack models: To mimic a realistic black-box setting, we pick a variety of object detectors, including two-stage detectors: Faster RCNN [396] and Libra R-CNN [370]; one-stage detector: RetinaNet [301]; and anchor-free detector: FoveaBox [257]. We use implementations of the aforementioned models from the `MMDetection` [90] code repository. The models in `MMDetection` are trained on `coco2017train`; therefore, while testing the detectors on VOC images, we only return the object labels available in VOC. The models under such adaptation still get good detection performance, as shown in tab:map.

Zero-Query attacks (ZQA and ZQA-PSPM): We evaluate two variants of our zero-query attack. The first variant (ZQA) ensures that the attack plan is chosen at random from the available set of context-consistent attacks and is determined using eq:osa. The second variant (ZQA-PSPM)

Table 5.1: Mean average precision (mAP) at IOU (intersection over union) threshold 0.5 of different detectors used in our experiments. Models are evaluated on VOC07 test set. **Legend:** Faster R-CNN (FRCNN), RetinaNet (Retina), Libra R-CNN (Libra), FoveaBox (Fovea).

Model	FRCNN	Retina	Libra	Fovea
mAP@.50	78.30%	78.51%	79.01%	77.68%

generates a context-consistent attack plan based on the PSPM matrix that was pre-computed for the given classification model, perturbation budget and evasion attack algorithm. See eq:osa-pspm. ZQA-PSPM is the key contribution of this paper, and our results demonstrate that that ZQA-PSPM provides better fooling success rate compared to ZQA and baselines.

Baselines and comparisons: We compare the ZQA and ZQA-PSPM schemes against two relevant baselines. The first is the context-agnostic zero-query attack, which we call "Context-Agnostic". In this attack, the attack plan that drives the scene perturbation is chosen randomly (i.e., without explicitly enforcing co-occurrence-based context). This means that some attack plans in the Context-Agnostic scheme may be context-consistent by accident, while others are context-inconsistent. Comparison against Context-Agnostic is performed with the aim of investigating the benefits of exploiting context while designing the attack plan.

We also compare with a second, more powerful baseline, which we refer to as the "Few-Query" approach [66]. In this scheme, the attacker is equipped with the co-occurrence matrix $\mathbf{G}$ but doesn't have the PSPM matrix. More importantly, the few-query attacker can query the victim system to find out whether the attack succeeded. Because of this, they don't need to perturb all the objects in the scene in one step. The few-query attacker proceeds as described in Algorithm 4. The

few-query attack is denoted as "Few-Query q" in tab:compare-frcnn and tab:compare-libra where q is the number of previous queries that the current attack is built on, $q \in \{0, 1, 2, 3, 4, 5\}$, thus "Few-Query 0" is identical to ZQA in terms of queries.

Attack generation: We use the PGD-based method to generate a perturbation on the whole image (as discussed in eq:PGD). We experiment with L_∞ perturbation budget $\epsilon \in \{10, 20, 30, 40, 50\}$. The step size $\lambda = 2$, and maximum number of iterations is 50. We observe that when an object is very close to or overlaps with the victim object, perturbing that object to any label different from the victim's target label reduces the success rate; thus, we map all objects whose regions have IOU ≥ 0.3 with the victim object to the victim's target label.

Evaluation metrics: We use the metric "context-consistent attack success rate" (or fooling rate) to evaluate the attack performance on a victim object detector. For an attack to be regarded as a successful context-consistent attack, it must (1) successfully perturb the victim object to the target label, and (2) pass the context-consistency check described in fig:consistency-check. We define the fooling rate as the percentage of the number of test cases for which the above two conditions are satisfied.

5.4.1 Experimental results on VOC dataset

Comparing zero-query attacks with few-query attacks: The fooling rates of the few-query attacks with different numbers of queries (Few-Query 0 to Few-Query 5) and the zero-query attack under white-box and black-box settings at different perturbation budgets are shown in tab:compare-

frcnn and tab:compare-libra. The settings for both tables are detailed in the captions. The fooling rates for the few-query attacks are cumulative; that is, the values reported for few-query-k accounts for successful attacks with $0, 1, \ldots, k$ queries.

We observe that ZQA can achieve higher fooling rates than the few-query attack for up to 4 queries in the white-box setting and up to 2 queries in the black-box setting. When PSPM is used to refine the zero-query attack plan (ZQA-PSPM), the fooling rate increases, outperforming up to 5 queries of the few-query attack in the white-box setting and up to 3 queries in the black-box setting. These results are consistent across several detector models tested. As ϵ reduces, the perturbation is not always enough to carry out the evasion attacks, and thus the fooling rates fall from $\epsilon = 50$ to $\epsilon = 10$.

The results clearly demonstrate the advantage of simultaneous context-aware perturbation of all objects in the scene. In many cases, using the PSPM to refine the context-aware attack further improves the fooling rate. While the few-query approach eventually outperforms the ZQA attack, recall that the former requires the attacker to communicate with the detector, which is either not always possible, or might expose the attacker.

5.4.2 Experimental results on COCO dataset

In this section, we repeat the object detection evaluation experiments for the COCO dataset. The models obtained from `MMDetection` are well trained on COCO2017 training set, and the evaluation results on COCO2017 validation set can be found in tab:map-coco. While the Mean Average Precision (mAP) scores are much lower than those observed for the VOC dataset (Table 5.1), these values are similar to the officially reported numbers in `MMDetection` repository. This confirms that the object detection algorithm for the COCO dataset – a more challenging dataset

than VOC – performs at a level close to the state of the art. The comparison of ZQA and ZQA-PSPM acting on the COCO dataset against our two baseline schemes is shown in tab:compare-frcnn-coco. As for the VOC dataset, the ZQA attack for the COCO dataset outperforms up to 3 attempts of the Few-Query attack (2 rounds of feedback) in the black-box transfer attack setting.

5.4.3 Evading context-agnostic defense

We tested against the commonly used context-agnostic JPEG defense and found that our attack is resilient. Our attack can still outperform up to 5 rounds of few-query attacks under the JPEG compression quality of 95, as shown in tab:compare-frcnn-jpeg95, corresponding to the setting in tab:compare-frcnn.

5.4.4 Visualization of sample images

In this section, we provide visual examples of scenes before and after perturbation. In doing so, we compare the zero-query scheme, the context-agnostic attack, and the few-query scheme that we developed to benchmark performance. All the results are for a transfer setting, i.e., the attacker creates the perturbations on a surrogate model which is different from the classification model used by the victim system. All the images are generated for the case in which the attacker's perturbation is made using a Faster R-CNN network, while the victim system system uses a RetinaNet model. The perturbation budget used to implement the evasion attack is $\epsilon = 10$.

fig:supp-fig1 provides an example in which the context-agnostic attack successfully perturbs the individual objects: chair $\rightarrow$ dog, chair $\rightarrow$ bus and chair $\rightarrow$ bird. However, the resulting list

of detected objects (dog, bus, bird) is context-inconsistent according to the co-occurrence matrix. Thus, the attack is detected. In contrast, the ZQA attack perturbs the objects as follows: chair → dog, second chair → second dog, dining table → person. The list of detected objects (dog, dog, person) is context-consistent, which fools the detector. This shows the basic use case of our context-aware approach.

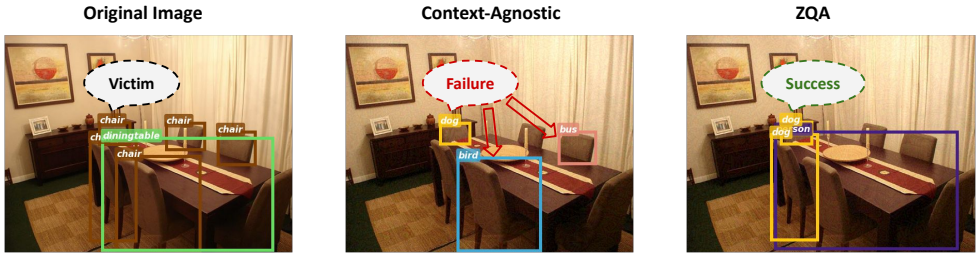


Figure 5.5: Detections on one original image and images perturbed by the context-agnostic attack and ZQA attack. The goal is to perturb the victim object, which is a chair on the top-left, to a dog. In the transfer attack, both the context-agnostic attack and ZQA attack successfully perturbs the chair to dog, along with some perturbations of surrounding objects. Even though context-agnostic attack is successful in perturbing victim to target, the attack still fails because the surrounding objects (bus and bird) are not context consistent according to the co-occurrence graph.

fig:supp-fig2 provides an example in which the few-query attack has perturbed the main victim object (sofa → bicycle), as well as one other helper object (chair → bicycle) in the scene. However, the attack fails because the victim system's detector does not detect the main victim object and relegates it to the background. In contrast, the ZQA attack, with the help of the perturbation success probability matrix (PSPM), chooses object perturbations that are most likely to succeed in a single attempt, i.e., sofa → bicycle, chair → person, and leaves the TV monitor unchanged. The

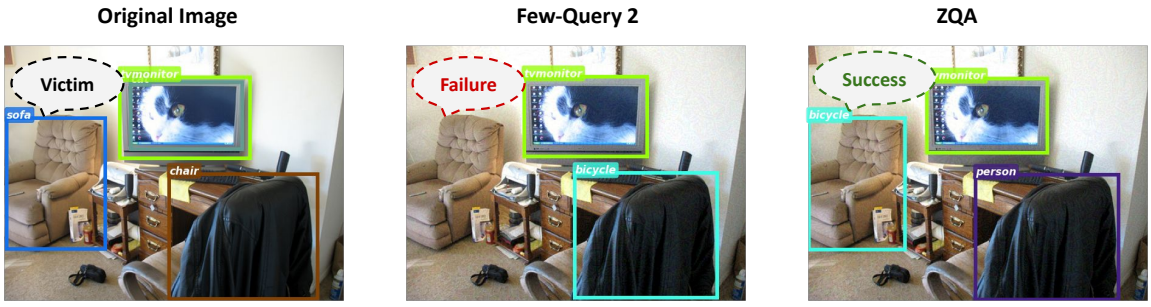


Figure 5.6: Detections on one original image and images perturbed by the few-query attack and the ZQA attack. The goal is to perturb the victim sofa to a target bicycle. Few-Query attack, building on 2 previous queries, perturbs the sofa to bicycle and the chair to bicycle as well. The TV monitor is not perturbed as it is context consistent. However, the attack failed to transfer to the victim model, in face, not detecting the sofa as a foreground object. Thus, the few-query attack fails. The ZQA attack additionally perturbs the chair to person. Since bicycle, person and TV monitor are all detected and are context-consistent, the attack successfully transfers.

perturbation applied to the sofa object is sufficient for it to be detected and misclassified as a bicycle. This attack is context-consistent by construction, and successfully fools the detector. We remark here that the vanishing effect scene above is not unique to the few-query attack. Indeed, evasion attacks which involve perturbing the entire scene while attempting to attack individual objects in the scene are susceptible to the vanishing effect. This occurs when the scene perturbation, constrained by the budget ϵ, is such that it causes one or more objects in the scene to not be detected. As expected, we observe this effect more often at lower perturbation budgets.

fig:supp-fig3 shows that, given more rounds of feedback, the few-query detector eventually gets enough information about the detector's decisions, and is able to perturb a large number of objects, thereby fooling the detector. The attack attempts to make the following changes: dog $\rightarrow$

boat, sofa $\rightarrow$ boat, cat $\rightarrow$ boat, person $\rightarrow$ boat. The victim system misclassifies the dog and the sofa as boats, but does not detect the person and the cat. Even with the vanishing artifact, we deem the few-query attack successful because it has successfully perturbed the victim object (dog $\rightarrow$ boat) and it has ensured that the detected objects form a context-consistent list. On the other hand, the ZQA attack intends to leave the person unchanged, while changing the other objects to boats. This attack fails because, at the given perturbation level $\epsilon = 10$, the attack left the person unchanged, altered the sofa to the boat, but caused the cat and the dog vanish into the background. This is a failed attack because the main objective of misclassifying the victim object, i.e., dog $\rightarrow$ boat, was not fulfilled. This shows that the few-query approach – given multiple attempts to enhance the attack – will eventually overwhelm the proposed ZQA attack which is only allowed a single attempt. One disadvantage of the few query-attack, as noted earlier, is that it requires access to the victim system's communication, thus exposing the attacker to the risk of being discovered. The ZQA attack does not have this limitation.

fig:supp-fig4 shows one of the failure modes of our approach. (This type of failure is also observed in general perturbation bounded evasion attacks, and in our case, it is also seen in some cases of the few-query attack, and the context-agnostic attack). The goal of the attacker is to perturb the horse to a cat. However, the attack made with the surrogate model does not correctly transfer to the black-box victim model. The detector recognizes the horse as a sheep, which is unintended for our targeted attack.

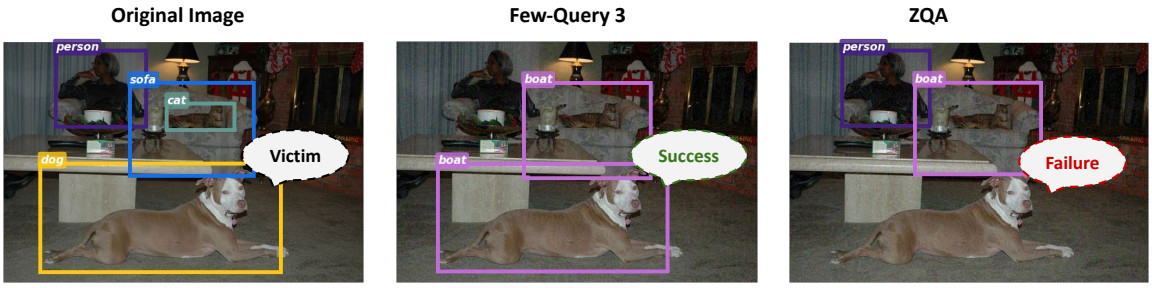


Figure 5.7: Detections on one original image and images perturbed by few-query attack and ZQA attack. The goal is to perturb the dog to a boat. The few-query attack, building on 3 previous queries, perturbs two objects to boats, and causes the person and the cat to vanish. The result is context-consistent and meets the desired goal. On the other hand, the ZQA attack leaves the person unchanged, perturbs the sofa to a boat, but causes the intended victim object (dog) and another object (cat) to vanish. Even though person and boat are context-consistent in the perturbed scene, the ZQA attack has failed because the intended victim object has vanished.

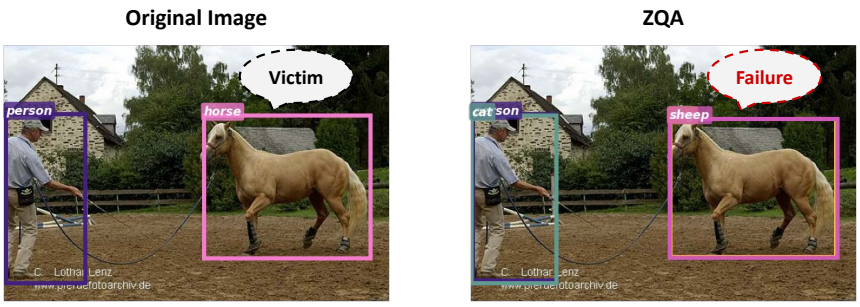


Figure 5.8: A failure case of ZQA attack. We observe that the perturbation of the victim object (horse $\rightarrow$ cat) does not succeed. Instead, the victim model classifies the perturbed horse as a sheep.

5.5 Discussion

Limitations and Future Work: We have assumed that the data distribution at the victim system is known. We also assume that the context is consistent across surrogate and victim systems. In practice, this is rarely the case. The attacker and the victim system may have different data distributions, overlapping but non-identical label sets, and thus similar but non-identical context models. A useful avenue for future work is to introduce controlled discrepancies between the distributions, context models, and label sets and examine their effect on the fooling rate of the ZQA attack.

While co-occurrence is a fundamental notion of context, it does not capture key properties such as relative size and location of the objects or the relationship between the object and the background. Extending the ZQA attack to more sophisticated context models is a topic for future research. Furthermore, evaluating the ZQA in situations where the attacker uses a different notion of context than the victim system – e.g., attacker uses semantic context, detector uses context learned from pixels – would help researchers understand the broader applicability of this work.

Another limitation is that it is expensive to precompute the PSPM. Unlike the context graph, which only depends on the dataset, PSPM is a function of the dataset, attack model and perturbation level. We need to measure the attack success rate for each surrogate model at each perturbation level, and for all possible perturbations of a given object.

Potential negative societal impact and mitigation: This paper, as any other work that investigates an attack method, may be used maliciously to generate attacks against victim systems. Our goal with this work is to reduce technical surprise, and to fuel the development of defenses against powerful attacks. This work already makes the case for using a context-aware detector to thwart sim-

ple attacks. To thwart ZQA attacks described in this paper, the detector can attempt to constantly update its training data and label sets, and to develop increasingly sophisticated context models.

5.6 Conclusions

In this paper, we craft a novel zero-query attack by exploiting "a context graph" that captures co-occurrence relations of objects in a natural image. Against context-aware detectors, the fooling rate is significantly higher than that achieved by a context-agnostic attack. Unlike prior query-based attacks, our attack is extremely hard to detect since it hinges on using a single attempt. It achieves fairly good fooling rates by choosing an attack plan (i.e., perturbing multiple objects simultaneously to ensure context consistency) which is likely to succeed. The key innovation is a PSPM that provides this information offline. We observe that the use of PSPM not only boosts fooling rates in white-box settings, but also carries over to the black-box setting (i.e., when the detector model is different from that of the attacker) consistently for different attacker-defender pairs.

Acknowledgments. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under agreement number HR00112090096. Approved for public release; distribution is unlimited.

Algorithm 3 Zero-Query Context-Consistent Attack Plan

Input:

A list of objects in the scene $\mathcal{A} = A_1 = \ell_v, A_2, \ldots, A_S$ where S denotes the total number of objects in the scene;

The desired target assignment consisting of (victim label $\ell_v \in \mathcal{A}$, target label ℓ_t) pair for one object;

Co-occurrence matrix $\mathbf{G}$ computed over the training data;

Co-occurrence threshold η;

Perturbation budget ϵ;

Perturbation success probability matrix $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}$ generated in advance from the training data

Output: Attack plan $\mathcal{B}$ that is context-consistent with $\mathbf{G}$ and most likely to succeed in a white-box setting

Initialize attack plan: $\mathcal{B} = \{\ell_t\}$

Obtain a set of labels that co-occur with ℓ_t, denoted as $\mathcal{T} \subset \mathcal{N}$ such that $\mathbf{G}(j, t) > \eta$ for all $\ell_j \in \mathcal{T}$ ▷ *find co-occurring labels*

Set running counter: $k = 2$

while $k \leq S$ **do**

  Compute $\ell^* = \underset{\ell_b \in \mathcal{T}}{\operatorname{argmax}} \mathbf{M}_{\mathcal{C},\epsilon,\alpha}(a, b)$, where a is the index of label A_k ▷ *find the best label to add*

  Update attack plan: $\mathcal{B} \leftarrow \mathcal{B} \cup \ell^*$ ▷ *add the selected label to the attack plan*

  Increment counter: $k \leftarrow k + 1$

end while

return $\mathcal{B}$

Algorithm 4 Few-Query Context-Consistent Attack Plans

Input:

A list of objects in the scene $\mathcal{A} = \{A_1 = \ell_v, A_2, \ldots, A_S\}$ where S denotes the number of objects in the scene;

The desired target assignment consisting of (victim label $\ell_v \in \mathcal{A}$, target label ℓ_t) pair for one object;

Co-occurrence matrix $\mathbf{G}$ computed over the training data;

Co-occurrence threshold η;

Number of victim system queries allowed $q \leq S$

Output: A sequence of attack plans $\mathcal{D}_k$ consistent with $\mathbf{G}$, where $k = 1, 2, ..., q$

Initialize attack plan: $\mathcal{D}_1 = \{\ell_t\}$

Obtain the label set $\mathcal{T} \subset \mathcal{N}$ such that $\mathbf{G}(j, t) > \eta$ for all $\ell_j \in \mathcal{T}$ ▷ *find co-occurring labels*

Set running counter: $k = 2$

while $k \leq q$ **do**

Uniformly sample from $\mathcal{T}$: $\widehat{\ell} \overset{iid}{\sim} U(\mathcal{T})$ ▷ *sample a label from $\mathcal{T}$*

Update attack plan: $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \cup \widehat{\ell}$ ▷ *add the sampled label to the attack plan*

Increment counter: $k \leftarrow k + 1$

end while

return $\mathcal{D}_k$

Table 5.2: Fooling rates (%) of different attack strategies under different L_∞ perturbation $\leq \epsilon \in \{50, 40, 30, 20, 10\}$. We compare ZQA and ZQA-PSPM with Context-Agnostic ZQA, and Few-Query attacks where feedback from blackbox (BB) models is allowed. The white-box (WB) is **Faster R-CNN** and three black-box models (BB1, BB2, BB3) are **RetinaNet, Libra R-CNN and FoveaBox** respectively. Fooling rate is counted as the percentage of attacks where victim is perturbed to a target label and all detected labels satisfy context consistency. Tested on 500 images from VOC 2007 test set which contain multiple (2-6) objects. Shaded cell indicates up to which few-query step, ZQA or ZQA-PSPM has better performance than few-query attack. Lighter shades are for ZQA, and darker shades are for ZQA-PSPM.

Method	$\epsilon = 50$				$\epsilon = 40$				$\epsilon = 30$				$\epsilon = 20$				$\epsilon = 10$			
	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3
Context-Agnostic	34.0	29.0	30.0	25.4	36.8	26.2	30.0	29.6	35.4	27.4	31.2	27.8	35.2	24.4	30.8	27.6	30.4	13.8	15.6	17.8
ZQA	90.0	46.6	52.2	54.0	92.0	48.0	55.0	51.8	91.6	46.0	57.0	52.2	87.4	39.6	50.4	51.0	65.2	21.0	23.8	24.2
ZQA-PSPM	**92.6**	**51.2**	**61.6**	**56.8**	**92.0**	**51.8**	**55.4**	**54.4**	**93.0**	**49.2**	**57.2**	**54.0**	**88.2**	**44.0**	**51.4**	**53.4**	**70.6**	**23.2**	**27.4**	**28.2**
Few-Query 0	60.0	29.8	29.8	34.8	64.2	34.6	34.2	39.8	66.2	34.2	35.0	37.8	61.2	29.2	30.8	35.8	48.2	14.8	14.0	20.2
Few-Query 1	64.4	35.8	40.4	43.0	68.0	41.0	44.2	49.4	69.6	41.8	45.2	47.6	68.2	36.0	40.2	43.4	58.4	21.6	23.8	27.8
Few-Query 2	77.6	48.0	56.2	59.0	80.0	50.0	52.6	57.2	78.4	47.0	54.2	54.6	77.6	43.8	50.0	49.8	70.4	27.0	29.6	34.2
Few-Query 3	86.8	55.4	65.0	65.8	89.6	55.4	58.6	62.0	86.2	54.8	62.0	61.2	86.4	49.6	57.2	55.4	77.0	31.4	34.0	39.8
Few-Query 4	91.6	60.0	71.8	69.4	95.2	61.4	63.8	66.0	91.2	58.0	68.2	67.2	89.6	53.2	60.6	59.8	81.4	34.2	37.8	43.2
Few-Query 5	95.0	61.8	75.0	73.4	97.2	62.8	68.0	69.4	96.2	61.2	71.0	70.6	93.0	56.6	65.2	63.4	85.2	35.8	40.2	46.2

Table 5.3: Follow the setting in tab:compare-frcnn but use **Libra R-CNN** as WB and use **Faster R-CNN, RetinaNet and FoveaBox** as BB1, BB2, BB3 respectively. Fooling rates (%) of different attack strategies under different L_∞ perturbation $\leq \epsilon \in \{50, 40, 30, 20, 10\}$ are as follows.

Method	$\epsilon = 50$				$\epsilon = 40$				$\epsilon = 30$				$\epsilon = 20$				$\epsilon = 10$			
	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3
Context-Agnostic	30.6	26.2	18.2	21.4	34.6	29.4	24.2	24.4	34.8	28.8	20.4	23.4	37.6	24.8	16.8	19.4	29.2	15.8	10.2	13.0
ZQA	90.4	47.8	33.0	43.4	91.8	48.4	32.2	40.0	88.6	48.6	33.0	41.8	**86.6**	39.4	25.6	35.0	64.4	23.8	13.4	21.6
ZQA-PSPM	**92.2**	**51.0**	**34.2**	**45.0**	**92.4**	**52.8**	**34.4**	**44.0**	**92.8**	**48.8**	**35.2**	**42.0**	86.2	**42.8**	**27.2**	**37.6**	**67.2**	**25.4**	**14.8**	**23.4**
Few-Query 0	63.0	35.6	25.8	34.0	64.2	37.0	27.0	32.2	63.2	37.0	25.2	33.0	62.2	33.2	23.2	32.2	44.0	19.0	11.0	19.6
Few-Query 1	66.8	43.6	32.4	41.0	68.2	43.6	32.4	39.2	67.8	43.8	32.0	39.4	68.0	42.8	28.8	38.2	58.6	28.0	17.8	27.4
Few-Query 2	77.4	51.8	38.2	47.8	76.8	51.4	37.6	47.2	78.2	51.4	37.4	45.6	76.6	47.4	33.2	43.0	67.6	33.8	21.4	33.0
Few-Query 3	87.8	57.8	45.0	55.2	87.4	59.4	43.0	54.0	85.4	56.4	42.8	50.4	85.8	53.0	36.8	50.2	75.4	36.4	24.0	34.6
Few-Query 4	93.4	61.0	47.4	59.4	93.4	63.6	47.0	58.8	91.4	60.6	45.4	55.8	92.4	57.4	39.8	54.0	80.8	40.2	25.8	37.6
Few-Query 5	96.8	64.8	49.6	61.8	97.0	67.0	49.6	61.0	95.8	63.2	47.0	58.0	95.2	59.6	41.8	57.0	84.6	41.6	27.2	39.6

Table 5.4: Mean average precision (mAP) at IOU (intersection over union) threshold 0.5 of different detectors used in our experiments. Models are evaluated on COCO2017 val set. **Legend:** Faster R-CNN (FRCNN), RetinaNet (Retina), Libra R-CNN (Libra), FoveaBox (Fovea).

Model	FRCNN	Retina	Libra	Fovea
mAP@.50	38.99%	35.13%	40.14%	45.78%

Table 5.5: Follow the setting in tab:compare-frcnn but use 500 images from COCO 2017 test set. Fooling rates (%) of different attack strategies under different L_∞ perturbation $\leq \epsilon \in \{50, 40, 30, 20, 10\}$ are as follows.

Method	$\epsilon = 50$				$\epsilon = 40$				$\epsilon = 30$				$\epsilon = 20$				$\epsilon = 10$			
	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3
Context-Agnostic	55.2	23.8	27.2	35.2	60.0	25.8	28.0	31.2	55.4	23.4	21.6	31.8	52.2	18.8	18.6	28.6	39.6	14.2	12.4	15.8
ZQA	82.2	29.8	35.4	43.6	82.8	30.2	35.8	43.2	81.0	**30.4**	31.0	40.0	76.0	23.8	26.4	**37.4**	52.0	14.2	15.6	21.8
ZQA-PSPM	**85.0**	**34.0**	**38.0**	**48.0**	**85.8**	**32.0**	**39.6**	**43.8**	**82.8**	29.8	**32.6**	**46.0**	**79.0**	**27.2**	**29.8**	36.8	**58.2**	**15.6**	**17.6**	**25.0**
Few-Query 0	71.4	27.0	24.0	37.8	73.4	25.6	24.0	35.0	68.4	21.8	18.2	34.0	63.8	21.2	17.6	28.4	47.2	13.2	9.6	18.4
Few-Query 1	80.0	34.2	34.2	46.6	80.0	33.6	34.2	44.0	79.0	29.8	27.2	44.4	72.6	26.0	24.6	36.8	56.4	19.2	15.4	26.0
Few-Query 2	83.4	37.8	41.4	51.4	84.0	39.4	39.0	50.4	84.2	34.2	33.2	49.8	79.2	31.4	30.2	43.6	62.4	21.8	19.0	30.6
Few-Query 3	86.2	40.6	46.2	55.2	86.8	41.8	42.2	54.6	86.2	36.6	39.2	53.6	81.6	33.2	34.8	46.6	66.8	23.6	21.2	34.0
Few-Query 4	88.0	42.8	48.0	57.8	89.8	42.8	45.4	56.6	87.8	37.8	42.0	55.4	84.4	35.2	38.0	49.2	69.0	24.2	23.0	36.2
Few-Query 5	89.2	45.0	52.4	59.2	92.0	44.4	48.0	57.8	89.8	39.6	45.0	57.6	85.8	36.0	40.6	51.6	71.4	25.6	25.4	38.2

Table 5.6: Follow the setting in tab:compare-frcnn but under the JPEG compression quality of 95. Fooling rates (%) of different attack strategies under different L_∞ perturbation $\leq \epsilon \in \{50, 40, 30, 20, 10\}$ are as follows.

Method	$\epsilon = 50$				$\epsilon = 40$				$\epsilon = 30$				$\epsilon = 20$				$\epsilon = 10$			
	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3
Context-Agnostic	34.6	26.4	30.0	25.6	33.4	23.6	25.0	26.0	34.4	26.4	28.8	27.0	38.4	24.0	23.6	25.8	28.2	9.4	11.0	14.8
ZQA	88.2	41.4	49.4	51.4	86.8	40.0	47.8	47.0	88.2	41.4	49.6	47.4	82.4	35.6	40.6	42.2	49.6	14.2	**16.8**	20.0
ZQA-PSPM	**89.2**	**42.8**	**50.2**	**53.8**	**90.2**	**41.2**	**48.6**	**49.8**	**92.8**	**44.2**	**52.2**	**51.2**	**83.6**	**36.4**	**42.0**	**44.2**	**55.8**	**15.6**	15.2	**21.4**
Few-Query 0	62.2	28.2	28.6	36.0	62.8	26.8	28.6	33.6	64.4	28.8	30.6	33.0	60.6	23.6	24.8	31.2	39.0	10.8	10.8	16.6
Few-Query 1	68.0	37.2	39.6	45.6	70.4	33.2	37.8	41.8	68.8	35.6	39.6	42.8	66.8	31.4	32.4	40.0	46.2	16.6	15.2	22.6
Few-Query 2	78.8	44.0	50.2	55.8	78.2	40.8	49.6	52.2	76.8	42.0	48.0	50.8	76.0	40.0	42.4	47.2	56.2	20.8	19.6	28.4
Few-Query 3	87.4	48.8	57.8	61.6	85.8	48.6	57.4	57.6	84.4	49.8	55.8	58.6	82.8	45.6	49.4	53.4	62.8	23.6	23.6	30.2
Few-Query 4	91.0	52.6	62.4	64.4	90.2	50.8	61.6	61.8	88.8	52.4	61.0	62.8	88.2	48.8	53.0	57.6	68.2	25.8	26.8	33.0
Few-Query 5	93.8	55.8	66.4	66.4	94.6	53.0	65.4	65.8	94.8	55.2	64.4	67.0	90.8	50.6	55.4	60.6	71.2	28.0	28.8	34.8

Chapter 6

Disguise without Disruption:

Utility-Preserving Face De-Identification

6.1 Introduction

Global privacy laws safeguard personal data, including regulations like GDPR [471] in Europe, HIPPA [1] and CCPA [2] in the US, and PIPL [3] in China. Particularly stringent for medical information and data from medical settings, these rules tightly control storage and distribution of patient images to ensure confidentiality. Yet, this data holds valuable potential, such as automating medical procedures and new AI-driven diagnoses. To tap into these datasets, scientists explore techniques for using sensitive images without compromising identity. Most methods focus on face obfuscation [362], blurring [154], pixelation [547], warping [258]), affecting image saliency. Face-swapping [212, 331, 93, 175, 72, 383, 8] is emerging as a promising solution.

Popularized through the notion of *deepfakes* [496], these deep-learning models are trained to replace any face in an image or video by another one (user-provided or AI-generated), while trying to preserve the overall saliency or specific facial attributes, such as perceived gender, expression, or hair color. While recent solutions can generate convincing results, they are not suitable for the targeted use cases as they lack formal *privacy* and *utility* guarantees for the resulting images. Face-swapping methods evade the confidentiality of the ID provider since the swapped face leaks the source ID. In addition, they lack proper mechanisms to maximize de-identification and minimize identity leakage of the target ID. Furthermore, they do not emphasize on maintaining the *utility* of resulting images, *i.e.*, they do not guarantee that the altered images can have the same function as the original ones for various downstream tasks. For example, a dataset would become *useless* for analysis if relevant non-biometric features are corrupted (*e.g.*, facial expressions have changed for sentiment analysis tasks) or for training recognition models if the altered images no longer match their annotations (*e.g.*, facial landmarks, gaze directions, head-pose orientations, *etc.*).

In this work, we aim to address the challenge of anonymizing images of individuals while ensuring privacy and maintaining high data utility. To this end, we propose *Disguise* (Deep Identity Swapper Guaranteeing Utility with Implicit Supervision from Experts), a de-identification method built upon face-swapping technology that offers formal guarantees regarding identity obfuscation and utility retention. Our main contributions are as follows:

- We propose a simple yet effective framework for face de-identification which generates natural faces with distinct identities from the original ones, while maintaining non-biometric attributes unchanged. unchanged.

- Unlike existing methods that pre-discard original face IDs, we condition the synthetic faces on the original ID vectors and maximize the distance to the original identities while ensuring differential privacy [138, 6], with randomization to prevent re-identification.
- We demonstrate superior results than state-of-the-art methods through extensive evaluation regarding the de-identification rate, utility preservation, and image quality of the resulting data over a large number of metrics.

6.2 Related work

Face Swapping. The topic of face swapping has received significant attention in research and is highly relevant, as evidenced by the large body of works dedicated to it [363, 280, 380, 93, 554, 515]. However, it presents inherent and important differences compared to face anonymization/de-identification. Face swapping aims to change the original identity to a specified target identity, whereas face anonymization shall not rely on actual identities, as it would otherwise compromise both target and source individuals. Moreover, the two domains consider different performance indicators and evaluation metrics. Anonymization aims at providing privacy-preserving guarantees, including face anonymization rate and non-re-identifiability [168, 303, 113, 452], which implies additional mechanisms compared to the face-swapping methods that prioritize preserving facial attributes while reckoning the visual quality of the injected identity [363, 515].

Face Anonymization. Although traditional methods such as blacking out, pixelation, and Gaussian blur [54, 170, 169, 361, 362] are effective in removing privacy-sensitive information, they drastically alter the original data distribution, resulting in a significant loss in *utility*. In other words, these methods generate anonymized images that are not suitable for downstream tasks such as gaze

estimation [248, 540], head-pose prediction [549, 190], facial-landmarks regression [121, 252], and expression estimation [495, 412] due to the lack of necessary visual information.

A significant amount of research on face anonymization approaches the problem as an image inpainting task, where the face region is first erased and then replaced with another. Early methods [168, 368] use a database of real faces to aggregate the new identity, while more recent methods [212, 331, 303] use generative models to synthesize fake identities based on the learned distribution. DeepPrivacy [212] is one of the pioneering works in this field, which reconstructs the missing face by taking the masked face and facial landmarks as inputs. However, the reconstructed face distribution suffers from bias as it is solely conditioned on its training data, leading to a tendency to generate smiling, young-looking faces. CIAGAN [331] is another work that uses facial masks and landmarks to generate new faces. However, it tends to generate faces with duplicated identities due to the length limitation of the one-hot vector. RePaint [324], a recent method based on diffusion models, generates photo-realistic faces with large facial variances, but it fails to maintain the utility of the faces and is sensitive to input distributions.

Some methods [175, 72, 383] have focused on making the anonymization process reversible, such as *Password* [175] and *RiDDLE* [276], which generate anonymized faces conditioned on a password that can be used to de-anonymize them. While such a feature can be desirable in some scenarios, it violates privacy regulations like GDPR [471] that protects *pseudonymous* data (data that has been de-identified from the data's subject but can be re-identified as needed by the use of additional information). In this work, we propose to anonymize faces in an irreversible manner. Other solutions [290, 87, 289, 303] incorporate notions of differential privacy [136, 138, 6] by adding adequately-calibrated random noise either at training or inference time, ensuring privacy lev-

els linked to their parameter ϵ. Or directly optimize in the latent space of StyleGAN [31]. However, they often neglect utility preservation (*e.g.*, they edit image background and utility attributes) and require complex post-processing, making them not readily applicable to anonymization tasks.

6.3 Methodology

In this section, we formalize our objectives, theoretically ground our work, and finally describe our proposed solution.

6.3.1 Problem Formulation

Privacy Utility Dual Optimization. Let $\mathcal{X} \subset \mathbb{R}^{3\times H\times W}$ be the image space, with $x \in \mathcal{X}$ an image depicting an individual. Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space, with $\mathcal{Z} \subset \mathbb{R}^{n_{\mathcal{Z}}}$ space of identity-distilled facial features (*i.e.*, facial features that uniquely identify an individual) and $d_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a distance function attached to space $\mathcal{Z}$. Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be another metric space, with $\mathcal{Y} \subset \mathbb{R}^{n_{\mathcal{Y}}}$ containing utility-distilled facial features (*i.e.*, features that are useful to downstream tasks) and $d_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a distance function relating to $\mathcal{Y}$. We note $f_{\mathcal{Z}} : \mathcal{X} \to \mathcal{Z}$ and $f_{\mathcal{Y}} : \mathcal{X} \to \mathcal{Y}$ the objective labeling functions respective to each domain.

We define a conditional generative function $G : \mathcal{X} \to \mathcal{X}$ parameterized by θ, that takes $x \in \mathcal{X}$ as input and returns an edited version $G(x) = \widetilde{x}$. Our goal is to learn a G such that *utility* is maximized (*i.e.*, $f_{\mathcal{Y}}(x) = f_{\mathcal{Y}}(\widetilde{x})$) and *privacy* is maximized (*i.e.*, $f_{\mathcal{Z}}(x)$ is distant from $f_{\mathcal{Z}}(\widetilde{x})$). In other terms, the output of G should contain the same utility attributes as the input and contain identity attributes different from the input beyond recognition. Formally, we want G to achieve Pareto optimality [415, 344] w.r.t. the aforementioned multiple objectives (*i.e.*, identity obfuscation

and utility preservation), accounting for their possible competition (depending on downstream tasks, utility and identity attributes may overlap), thus minimizing the following objective:

$$\min_{\theta}\Big(-\mathbb{E}_{x\in\mathcal{X}}\left[d_{\mathcal{Z}}\big(f_{\mathcal{Z}}\left(x\right),f_{\mathcal{Z}}\circ G_{\theta}\left(x\right)\big)\right],\\ \mathbb{E}_{x\in\mathcal{X}}\left[d_{\mathcal{Y}}\big(f_{\mathcal{Y}}\left(x\right),f_{\mathcal{Y}}\circ G_{\theta}\left(x\right)\big)\right]\Big)^{\intercal} \tag{6.1}$$

Before tackling the challenges of multi-objective optimization that such a task brings, one has to consider how to model the unknown objective distance and labeling functions $d_{\mathcal{Z}}, f_{\mathcal{Z}}$ and $d_{\mathcal{Y}}, f_{\mathcal{Y}}$ for the identity and utility space respectively. We argue that identity and utility are conceptually subjective, *i.e.*, different authoritative entities have different definitions and target features assigned to each concept. *e.g.*, given a picture of a person, each human or algorithmic agent will rely on different features (facial landmarks, eye color, *etc.*) and their own subjective judgment to certify the person's identity, as there is no absolute objective function to perform the ill-posed mapping of a facial picture to an identity. Similarly, the concept of *utility* is conditioned by a set of target tasks or the agents in charge of said tasks. *e.g.*, an image with the person's face completely blurred could still be *used* by a person-detection algorithm, but would be *useless* for facial landmark regression.

Therefore, we propose to rely on predefined agents (*experts*) to provide the identity and utility definitions to guide the optimization of our model [168]. We thus consider some parameterized models $h_{\mathcal{Z}}$ and $h_{\mathcal{Y}}$ pre-optimized to approximate their respective objective labeling functions $f_{\mathcal{Z}}$ and $f_{\mathcal{Y}}$. Note that we make no assumption on the architecture or training of each of these models (we demonstrate with various state-of-the-art identity extraction and recognition models). Without loss of generality and to account for individual bias, we define $H_{\mathcal{Z}} = \left\{h^i_{\mathcal{Z}}\right\}_{i=1}^{k_{\mathcal{Z}}}$ and $H_{\mathcal{Y}} = \left\{h^i_{\mathcal{Y}}\right\}_{i=1}^{k_{\mathcal{Y}}}$ as sets of $k_{\mathcal{Z}}$ and $k_{\mathcal{Y}}$ unique models which differ in terms of architecture and/or training regime, *c.f.* mixture-of-experts theory [342, 330, 116]. We demonstrate in this paper how these identifica-

tion/utilization experts can be leveraged in an adversarial/collaborative framework to train g towards a satisfying optimum.

Identity Obfuscation Guarantees. To provide formal de-identification guarantees, we ground our work in the extensive theory on ϵ-differential privacy (ϵ-DP) and ϵ-local-differential privacy (ϵ-LDP, relevant when obfuscation should be performed without global knowledge) applied to identity-swapping functions [136, 138, 6, 526, 303, 113, 452, 384]. Let $\psi : \mathcal{Z} \rightarrow \mathcal{Z}$ be a function that performs ID obfuscation, *i.e.*, taking an identity vector z and returning a new one $\widetilde{z}$ that maximizes $d_{\mathcal{Z}}(z, \widetilde{z})$. We consider that an approximate but randomized function $\psi^{\epsilon} : \mathcal{Z} \rightarrow \mathcal{Z}$ satisfies ϵ-LDP if, for any two adjacent inputs z, $z' \in \mathcal{Z}$ and for any subset of outputs $Z_s \subseteq \mathrm{range}(\psi^{\epsilon})$, it holds that $\mathrm{P}\left(\psi^{\epsilon}(z)\right) \leq e^{\epsilon}\mathrm{P}\left(\psi^{\epsilon}(z')\right)$. Given $\Delta\psi = \sup_{z,z' \in \mathcal{Z}} \|\psi(z) - \psi(z')\|_1$ the sensitivity of ψ, Laplace noise is commonly leveraged to define an ϵ-DP version of the function: $\psi^{\epsilon}(z) \triangleq \psi(z) + \left(\mathrm{Lap}\left(\Delta\psi/\epsilon\right)\right)^{n_{\mathcal{Z}}}$ [136, 138, 6, 303]. We demonstrate that to ensure ϵ-LDP, the $d_{\mathcal{Z}}$-maximization property of the identity-obfuscation function has to be relaxed. The manifold of identity vectors generated by an identification function $h_{\mathcal{Z}}$ is bounded by the range of said function. In such a space and for any Euclidean distance $d_{\mathcal{Z}}$, a non-relaxed version of ψ would be the bijective (and thus non-private) function ψ_{opp} mapping an ID vector to its opposite. No other function (*e.g.*, ψ^{ϵ}) could ensure $d_{\mathcal{Z}}$-maximization. Therefore, in this work, we consider the inherent trade-off between maximizing swapping-based identity obfuscation and ensuring differential privacy, and we propose a variety of solutions ψ^{ϵ} tailored to different needs (as illustrated in Figure 6.3, and more details in Proposed Solution Section).

Non Re-identifiability. Another important aspect to consider in privacy-preserving applications is *non-invertibility*. If the de-identified data can be re-identified with additional information, then the

operation is not truly anonymization but *pseudonymization*. For example, with the correct password for Password [175] and RiDDLE [276], or using the opposite ID for ψ_{opp}, the original ID is compromized. We empirically demonstrate that the proposed obfuscation solutions achieve varying degrees of robustness to such re-identification efforts.

In the remaining of the section, we explain how we define and train g to ensure privacy-preserving non-invertible identity swapping in images and utility preservation.

6.3.2 Proposed Solution

The proposed architecture can be defined as the composition of a face-swapping model $g : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{X}$, an identity extractor $h_{\mathcal{Z}} : \mathcal{X} \rightarrow \mathcal{Z}$, and an identity obfuscation function $\psi^{\epsilon} : \mathcal{Z} \rightarrow \mathcal{Z}$, such that $G(x) = g\left(x, \psi^{\epsilon} \circ h_{\mathcal{Z}}(x)\right)$. Given a facial image x, $h_{\mathcal{Z}}$ extracts the vector z encoding the identity of the depicted person. This vector z is passed to the privacy-enabling function ψ^{ϵ}, which returns a synthetic identity $\widetilde{z}$ that maximizes obfuscation. Finally, the face-swapper model g edits the original image x to inject the fake identity $\widetilde{z}$, resulting in an image $\widetilde{x}$ where the original visual identifying attributes are replaced by those encoded in $\widetilde{z}$. Additionally, during its training, g relies on the feedback of tasks-specific models $h^i_{\mathcal{Y}} : \mathcal{X} \rightarrow \mathcal{Y}$ to ensure that the utility of $\widetilde{x}$ is maintained compared to x. We expand on each block in the following paragraphs.

Identity Extraction. As mentioned in Section 6.3.1, we propose to extract the identity information from facial images via model ensembling [342, 330, 116], to ensure generalizability as well as to limit the impact of models' bias (as we assume no control over the architecture or training regimen of selected identity-expert models). Therefore, given a set $H_{\mathcal{Z}} = \left\{h^i_{\mathcal{Z}}\right\}_{i=1}^{k_{\mathcal{Z}}}$ of ID extractors, we define $h_{\mathcal{Z}}$ as the ensemble method $h_{\mathcal{Z}}(x) = \text{MLP}_{\theta_z}\left(\|_{i=1}^{k_{\mathcal{Z}}} h^i_{\mathcal{Z}}(x)\right)$, *i.e.*, concatenating (symbol $\|$)

the $k_{\mathcal{Z}}$ predicted vectors together and merging them into $z \in \mathcal{Z}$ via a multilayer perceptron (MLP) with parameters θ_z and $\tanh$ as final activation.

Identity Transformation. A variety of techniques can be considered to perform the identity transformation ψ, as shown in Figure 6.3. If we were to maximize the distance between the original and obfuscating IDs, the optimal function would be $\psi_{\text{opp}}(z) = -z$ since $\arg\max_{z'} d_{\mathcal{Z}}(z, z') = -z$ in our normalized Euclidean identity space. However, such a function is reversible, making it easy to re-identify the original individual by taking the opposite of the pseudonymized ID. A more secure solution would be a parametric function, *e.g.*, $\psi_{\text{mlp}}(z) = \text{MLP}_{\theta_\psi}(z)$, trained to optimally fool $h_{\mathcal{Z}}$. As a non-explicit function, ψ_{mlp} is more challenging to invert, though not impossible with the access to the model or its parameters θ_ψ (*c.f.* gradient-based attacks [153, 486]). To increase robustness and ensure ϵ-LDP, we can add dimension-wise noise to the inner operation, *i.e.*, $\psi^{\epsilon}_{\text{mlp}}(z) = \text{MLP}_{\theta_\psi}\left(z + (\text{Lap}\,(\beta))^{n_{\mathcal{Z}}}\right)$, with $\beta = \frac{\Delta\psi_{\text{mlp}}}{\epsilon}$. The larger β is set (*i.e.*, the smaller ϵ is), the more noise is applied to the original ID vector before further MLP-based obfuscation. Therefore, larger β provides stricter privacy guarantee and robustness but adversarial affects the ability of $\psi^{\epsilon}_{\text{mlp}}$ to learn how to fool identification experts $H_{\mathcal{Z}}$.

To better navigate this trade-off and guarantee a more continuous space for the noise application, we leverage the properties inherent to variational autoencoders (VAEs) [254]. We introduce a variational encoder-decoder (VED) to transform the identity vector, *i.e.*, $\psi^{\epsilon}_{\text{ved}}(z) = \text{VED}_{\theta_{\psi'}}(z)$. This model's encoder predicts the parameters (μ, σ) of the latent data distribution (assumed to be Gaussian). A latent vector v_z is picked as $\mu + \sigma\eta$ with $\eta \sim (\mathcal{N}(0,1))^{n_v}$ (*c.f.* reparameterization [255]) then passed to the decoder. While a VAE decoder would reconstruct the input identity from v_z, our VED decoder should generate a new, distant identity. During inference, we sample v_z as

$\mu + \sigma\left(\mathrm{Lap}\left(\alpha\right)\right)^{n_v}$ to meet ϵ-LDP, with n_v dimension of latent space and $\alpha = \frac{\Delta\psi_{\text{ved}}}{\epsilon}$. To train either of these models, we enforce cosine dissimilarity between the original and generated ID vectors:

$$\mathcal{L}_{\text{deid}} = 1 + \frac{z \cdot \widetilde{z}}{\|z\|_2 \|\widetilde{z}\|_2}. \tag{6.2}$$

For the VED model, we add to this criterion the usual Kullback–Leibler divergence (KLD) loss $\mathcal{L}_{\text{kld}}$ [255, 254].

Face Swapping. Once the fake identity vector $\widetilde{z}$ is generated, it is passed to the face-swapping model g, along with the original image x. Similar to existing solutions [93, 380, 303], g is composed of three modules: (1) an image encoder that extracts identity-unrelated features ν; (2) an ID injector that aggregates ν and $\widetilde{z}$ into a vector encoding the content of the obfuscated image $\widetilde{x}$; (3) a decoder conditioned on this vector that generates $\widetilde{x}$. These existing works also share similar losses that we borrow and adapt:

$$\begin{aligned}
\mathcal{L}_{\text{mix}} &= \|g(x, \widetilde{z}) - g(x, z)\|_1; \\
\mathcal{L}_{\text{gen}} &= \sum_{i=1}^{k_d} \log\left(1 - D_i(x, \widetilde{x})\right); \\
\mathcal{L}_{\text{id}} &= \sum_{\hat{z} \in \{z, \widetilde{z}\}} \Big(1 - \frac{\hat{z} \cdot \hat{z}_h}{\|\hat{z}\|_2 \|\hat{z}_h\|_2}\Big);
\end{aligned} \tag{6.3}$$

with $\hat{z}_h = h_{\mathcal{Z}}\left(g(x, \hat{z})\right)$. Here, $\mathcal{L}_{\text{mix}}$ is a mixing loss to ensure implicit disentanglement of ID features (encoded in z or $\widetilde{z}$) and residual features (*i.e.*, ν). $\mathcal{L}_{\text{gen}}$ pits the generator against k_d discriminators D to ensure realistic results preserving image saliency, *c.f.* recent GAN solutions [482, 212, 93] (we also use their weak-feature matching loss, further ensuring the high-level semantic alignment between the image pairs). Finally, $\mathcal{L}_{\text{id}}$ enforces cosine similarity between the injected identity $\hat{z}$ and the one observed by the identification model $h_{\mathcal{Z}}$ in the resulting image. Combined together, along

with $\mathcal{L}_{\text{deid}}$ and $\mathcal{L}_{\text{kld}}$ (using weighting hyperparameters), these losses form the overall objective for our privacy-enforcing face-swapping solution G.

Utility Preservation. Existing face-swapping methods [212, 93, 380, 303] claim that their adversarial and feature-matching losses ensure the preservation of non-identifying content. However, such supervision is too weak to guarantee that the images will maintain their utility w.r.t. downstream tasks, especially for tasks relying on small attention regions (*e.g.*, gaze estimation). We thus complement the aforementioned objective with a criterion that leverages the implicit expertise of tasks-relevant models $H_{\mathcal{Y}}$:

$$\mathcal{L}_{\text{uti}} = \sum_{i=1}^{k_{\mathcal{Y}}} \lambda_{\text{uti},i} \| h_{\mathcal{Y}}^{i,l}(x) - h_{\mathcal{Y}}^{i,l}(\widetilde{x}) \|_1, \tag{6.4}$$

with $h_{\mathcal{Y}}^{i,l}(\cdot)$ the features returned by the last differential non-softmax layer l of model $h_{\mathcal{Y}}^{i}$, and $\lambda_{\text{uti}} \in \mathbb{R}^{k_{\mathcal{Y}}}$ hyperparameters weighting the task/expert contributions. Hence, $\mathcal{L}_{\text{uti}}$ imposes that altered images contains the same utility attributes as original images, as expected by tasks-relevant models.

Note that the entire solution $G(x) = g\left(x, \psi^{\epsilon} \circ h_{\mathcal{Z}}(x)\right)$ is end-to-end differentiable, thus single-pass trainable. In practice, we leverage its modularity and train each component separately before jointly fine-tuning. Scalar hyperparameters weigh the contribution of each loss to the total objective (we fix $\{\lambda_{\text{id}}, \lambda_{\text{deid}}, \lambda_{\text{mix}}, \lambda_{\text{uti,eye}}, \lambda_{\text{uti,emo}}, \lambda_{\text{kld}}\} = \{30, 30, 10, 2, 2, 0.2\}$).

6.4 Experiments

We now describe our experimental setup and compare with other methods in terms of privacy robustness and data usability. More details in supplementary material.

6.4.1 Experimental Protocol

Datasets. We use multiple datasets for taining and evaluation. We train our models on VGGFace2 dataset [74], which totals 3.31 million images with 9,131 identities. We use multiple datasets for evaluation, including LFW [203] (13,233 face images and 5,749 identities) for utility and de-identification performance, CelebA-HQ [242] (30,000 face images) for utility evaluation, and WFLW [499] (10,000 face images) for the training usability w.r.t. the downstream task of landmark detection.

Identity and Utility Models. To demonstrate the genericity of our method, we consider a variety of pretrained face-identification networksand of utility networks over different recognition tasks. As identity experts, we use ArcFace [122], AdaFace [251], FaceNet [413], and SphereFace [306]. Either ArcFace ($h_{\mathcal{Z}}^{\text{arc}}$), AdaFace ($h_{\mathcal{Z}}^{\text{ada}}$), or both ($h_{\mathcal{Z}}^{\text{mix}}$) are used to guide g during training (*c.f.* Equation 6.4); FaceNet and SphereFace are used only for evaluation. For the downstream tasks, we use ETH-XGaze [540] (noted $h_{\mathcal{Y}}^{\text{eye}}$) for gaze estimation, DAN [495] ($h_{\mathcal{Y}}^{\text{emo}}$) for facial expression recognition, or both ($h_{\mathcal{Y}}^{\text{mix}}$) to provide utility feedback during training. During evaluation, we use L2CS-Net [7] for gaze estimation, DeepFace (DF) [416] for emotion recognition, and RetinaFace [121] and Dlib [252] for landmark detection.

Metrics. We employ the commonly-used validation rate and verification accuracy as metrics for evaluating privacy preservability [413, 306, 122, 251]. The validation rate is defined as the true

positive rate (TPR) at certain false positive rate (FPR), *e.g.*, TPR @ FPR=1e-3. Verification accuracy is the percentage of image pairs correctly classified as the same/different person using the best ℓ_2 distance threshold. The verification accuracy of random guessing is thus 50%, which is what anonymization aims at. To measure utility preservation, we use ℓ_2 pixel distance and normalized mean error (NME) for facial landmark detection, mean absolute error (MAE) for gaze estimation, and accuracy for emotion recognition. For image quality, we use SER-FIQ [448].

Comparison. We consider various de-identification methods, including DeepPrivacy [212], DeepPrivacy2 [211], CIAGAN [331], Password [175], and RePaint [324]. For readibility of the tables, we denote different versions as "Ours (a, b, c)" where a fixes the identity model(s) $h^a_{\mathcal{Z}}$ used, b the transformation function ψ^ϵ_b, and c the utility model(s) $h^c_{\mathcal{Y}}$. For simplicity, we use "Ours (arc, ved, eye)" as our default method unless otherwise mentioned. We demonstrate the impact of different transformation models and identity/utility experts in ablation studies.

Architecture. We build our face-swapping model g based on the architecture and training framework proposed in SimSwap [93]. The identity-merging module uses a MLP_{θ_z} with two layers and feature sizes of [1024, 1024, 512]. The identity-transforming MLP_{θ_z}, on the other hand, is a 3-layer network with feature sizes [512, 2048, 1024, 512]. The VED encoder consists of two dense layers of sizes [512, 1024, 1024], followed by two parallel layers of size [512] to predict the mean and variance in latent space. The VED decoder has three dense layers of sizes [512, 1024, 1024, 512]. We use $\tanh$ as final activation throughout the model.

Training. We apply the Adam optimizer [253] with $\beta_1 = 0$ and $\beta_2 = 0.999$, learning rate 10^{-4}, and a batch size of 4. We train our pipeline in two stages: (1) we first pre-train the face-swapper

g according to [93] for 1M iterations; (2) then we fine-tune it together with utility module and ID transformer for another 100k iterations. This is illustrated in Figure 6.6.

Evaluation. For the de-correlation evaluation presented in Figure 6.5(e), the MLP attacker networks consist of three layers of feature sizes [512, 2048, 1024, 512], tasked to reconstruct the original identity embedding from the obfuscated one extracted from the edited image. We train one attacker specific to each obfuscation method (CIAGAN [331], DeepPrivacy [212], ours, *etc*.). We use Adam optimizer with a learning rate of 10^{-3} and a total epoch of 100 epochs. We trained the decoders on CelebA-HQ and evaluated them on LFW.

6.4.2 Privacy: Obfuscation Evaluation

De-identification performance. As shown in Table 6.1, we achieve near perfect de-id rate, *i.e*., with a validation rate close to 0 and verification accuracy close to 50%, outperforming other methods by a significant margin, and is even more secure than randomly picking replacement images from the dataset. Figure 6.5(b) presents the ℓ_2 distance histogram for original positive pairs, original negative pairs, and original-anonymized positive pairs on LFW [203], and Figure 6.5(c) shows the ROC curves of validation rate. We observe that *Disguise* creates image pairs that are close to the negative distribution, hence perfect obfuscation. We also achieve the highest facial image quality, see Figures 6.1 and 6.4 for visual reference. Among other comparing methods, it is worth noticing that Password fails to de-identify images, hence the highest validation rate. CIAGAN and RePaint are better than Password in de-identification, however they suffer from low facial image quality due to high artifacts and distortions.

Table 6.1: Identification / validation rate and image quality evaluation over edited LFW data.

Methods	TPR (%) @ FPR=10^{-3} / Accuracy (%) ↓				FIQ ↑
	FaceNet	SphereFace	AdaFace	Average	SER
Original	93.8 / 97.1	87.9 / 96.2	95.4 / 97.7	92.4 / 97.0	0.77
DeepPrivacy	7.3 / 73.8	2.9 / 70.9	4.6 / 68.6	4.9 / 71.1	0.67
DeepPrivacy2	1.7 / 62.5	1.0 / 61.5	2.2 / 62.2	1.6 / 62.1	0.68
CIAGAN	1.8 / 64.5	1.0 / 59.0	5.6 / 71.0	2.8 / 64.8	0.58
Password	31.7 / 79.1	17.1 / 73.5	51.0 / 84.0	33.3 / 78.9	0.69
RePaint	2.8 / 67.7	1.1 / 63.5	3.6 / 68.5	2.5 / 66.6	0.54
Ours	**0.03 / 50.0**	**0.03 / 50.0**	**0.00 / 50.0**	**0.02 / 50.0**	**0.90**

Original and anonymized ID de-correlation. We consider scenarios where malicious attackers attempt to link anonymized IDs with their original IDs, allowing them to perform inversion inference on the anonymized IDs and recover the original ones. We use encoder-decoder networks to learn the correlation on existing original-anonymized image pairs. Figure 6.5(e) shows the results of using MLPs to decode obfuscated IDs from CelebA-HQ [242] while trained on LFW [203] using original IDs as supervision. While methods like DeepPrivacy, CIAGAN, and RePaint are inherently robust to inversion attacks since the original face region is entirely erased, and their networks are solely tasked with inpainting the blank region, our method still offers de-correlation on par with these methods, suggesting that our method is also resilient to inversion attacks.

Table 6.2: Utility performance comparison of different anonymization methods over diverse downstream tasks on LFW and CelebA-HQ datasets [203, 242].

Dataset	Methods	Facial landmarks (L2 pixel distance ↓)								Gaze estimation (MAE° ↓)				Emotion	
		RetinaFace (5 points)				Dlib (68 points)				L2CS-Net		ETH-XGaze		(Accuracy % ↑)	
		All	Eyes	Nose	Mouth	All	Eyes	Nose	Mouth	Pitch	Yaw	Pitch	Yaw	DAN	DF
LFW	DeepPrivacy	23.9	13.1	9.9	16.5	263.0	32.7	25.1	89.0	7.7	13.6	8.0	16.3	27.1	34.3
	DeepPrivacy2	31.2	18.4	14.4	19.6	385.7	59.9	49.9	120.6	9.2	12.2	7.8	15.1	22.4	30.2
	CIAGAN	14.6	9.3	5.5	9.2	348.2	59.0	31.2	97.7	8.8	14.6	7.8	16.9	32.5	36.9
	Password	17.4	10.4	7.7	11.1	204.8	**26.5**	**19.3**	**55.4**	10.5	24.7	7.7	11.5	45.9	43.4
	RePaint	66.1	30.8	32.2	47.3	1103.1	133.5	152.1	432.0	11.3	18.1	9.2	18.8	17.3	19.4
	Ours	**12.9**	**7.7**	**5.6**	**8.2**	**203.3**	28.8	19.8	60.0	**6.8**	**8.4**	**5.6**	**10.0**	**46.2**	**47.0**
CelebA-HQ	DeepPrivacy	13.1	4.8	4.3	10.8	293.9	30.4	24.2	113.0	7.0	8.7	6.7	10.1	41.0	45.5
	CIAGAN	14.9	10.3	4.6	8.6	365.6	79.2	35.6	91.5	8.7	13.7	7.5	13.4	38.0	44.4
	RePaint	9.9	**3.0**	4.6	7.5	249.1	**22.7**	29.7	95.9	6.2	8.0	5.5	8.0	50.4	55.8
	Ours	**6.7**	3.4	**3.3**	**4.3**	**196.0**	25.1	**20.0**	**59.9**	**5.6**	**6.2**	**4.6**	**5.9**	**61.9**	**59.6**

6.4.3 Utility: Usability Evaluation

Utility corruption in anonymized images. Our method demonstrates superior utility preservation compared to others across datasets (Table 6.2). We highlight our approach's excellence through qualitative comparison (Figure 6.1 and 6.4). DeepPrivacy lacks facial attribute preservation, exhibiting bias towards smiles and youth. CIAGAN bears heavy artifacts; Password yields blurry and easily re-identifiable outcomes. RePaint excels with in-distribution faces (RePaint is trained on CelebA-HQ thus has improved performance on the same dataset in Table 6.2), but it fails elsewhere and doesn't retain original attributes. For challenging scenarios, like heavy occlusion (*e.g.*, masks), CIAGAN and DeepPrivacy falter, unlike our effective face-swapping model.

Usability of anonymized images as training data. We have demonstrated utility attribute non-corruption by comparing performance of pretrained task-specific models on obfuscated versus orig-

Table 6.3: Usability of de-identified datasets for the training of task-specific models (facial landmark detection on WFLW).

Methods	Normalized Mean Error (NME) ↓						
	all	pose	illu	occ	blur	mu	exp
Original	.039	.068	.039	.047	.045	.038	.043
DeepPrivacy	.058	.100	.057	.072	.066	.060	.066
CIAGAN	.055	.087	.054	.064	.061	.053	.060
Ours	**.047**	**.079**	**.047**	**.056**	**.054**	**.046**	**.050**

inal data. Now, we advance toward the initial motivation of data anonymization for new solutions, evaluating how utility networks trained from scratch on anonymized data perform on real, unseen samples. Ideally, these privacy-preprocessed models should match performance of those trained on original, non-obfuscated data. Taking facial landmark detection as an example on the WFLW dataset [499] (98 landmarks per image), we split data into training/testing sets (7,500/2,500) and generate obfuscated training data using mentioned methods (test data remains unaltered). We use an HRNetv2-W18 model [480] for the task, trained for 60 epochs with Adam optimizer [253] ($\beta_1 = 0$, $\beta_2 = 0.999$), learning rate 10^{-4}, and batch size 64. Table 6.3 shows models on obfuscated data perform worse (higher NME of facial landmarks) than the one on original data. Our anonymized data model demonstrates the smallest accuracy drop, confirming higher utility preservation for downstream tasks while maintaining privacy.

Table 6.4: Re-identifiability of our ID transformation methods.

Methods	TPR (%) @ FPR=1e-3 ↓ (LFW data)			
	Swapped		Inverted	
	FaceNet	Sph.Face	FaceNet	Sph.Face
Ours (arc, opp, Ø)	0.63	0.03	67.03	53.07
Ours (arc, ved, emo)	0.23	**0.00**	**12.03**	7.10
Ours (arc, ved, eye)	0.03	0.03	13.03	**6.43**
Ours (arc, mlp, eye)	**0.00**	**0.00**	52.90	45.97
Ours (arc, mlp, emo)	0.03	**0.00**	49.77	45.97
Ours (arc, mlp, mix)	**0.00**	**0.00**	50.23	44.67
Ours (mix, mlp, eye)	0.07	**0.00**	36.70	34.70

6.4.4 Ablation Study

Here we demonstrate the impact of different transformation models, identity and utility experts. We also delve deep into assessing how diverse ID extraction methods, ID transformation techniques, and utility experts can collectively influence the overall obfuscation pipeline.

Impact of transformation models on re-identifiability. As justified in Section 6.3 and experimentally measured in Table 6.4, ψ_{opp} would suffer high re-identification, *i.e*. we can recover the original ID using the opposite of transformed ID. MLP-based transformations outperforms opposite transformation but VED-based transformations yield the best results in terms of de-identification and non-invertibility, confirming the superiority of our proposed solution.

Table 6.5: Effect of $\psi^{\epsilon}_{\cdot}$ noise w.r.t. (re-)identifiability.

Methods		TPR (%) @ FPR=1e-3 ↓ (LFW data)			
		Swapped		Inverted	
Network	Noise	FaceNet	Sph.Face	FaceNet	Sph.Face
MLP	$\beta = 0.0$	**0.00**	**0.00**	52.90	45.97
	$\beta = 0.5$	0.40	**0.00**	23.20	20.80
	$\beta = 0.9$	5.90	2.50	**3.07**	**1.30**
VED	$\alpha = 1.0$	**0.03**	0.03	13.03	6.43
	$\alpha = 2.0$	0.37	**0.00**	7.43	3.20
	$\alpha = 3.0$	0.37	0.10	**5.37**	**2.23**

The introduction of stochastic operations in alignment with ϵ-LDP further strengthen the solution. As shown in Table 6.5, the higher the amount of β or α noise introduced (*i.e.*, the lower ϵ), the more robust to attacks the method becomes, but the lower the original de-identification rate (the noisier the data, the harder it is to synthesize an ID that maximizes obfuscation). This negative impact is however better mitigated by the proposed VED. We provide further insights in supplementary material.

Efeects of using multiple ID extractors. As shown in Table 6.4, MLP-based transformations relying on multiple identity extractors, *i.e.*, "Ours (mix, mlp, eye)", perform better than versions with only one ID expert. We attribute the increased robustness to the combined knowledge of the two algorithms which capture more varied ID-related features that are then obfuscated.

Table 6.6: Identification / validation rate (↓, lower = better) and image quality evaluation (↑, higher = better) over edited LFW data.

Methods	TPR (%) @ FPR=10^{-3} / Accuracy (%) ↓				FIQ ↑
	FaceNet	SphereFace	AdaFace	Average	SER
Original	93.83 / 97.1	87.90 / 96.2	95.43 / 97.7	92.39 / 97.0	0.77
Ours (arc, opp, Ø)	0.63 / 51.5	0.03 / 50.0	0.03 / 50.0	0.23 / 50.5	0.81
Ours (arc, ved, emo)	0.23 / 50.1	**0.00** / 50.0	0.07 / 50.0	0.10 / 50.0	0.90
Ours (arc, ved, eye)	0.03 / 50.0	0.03 / 50.0	**0.00** / 50.0	0.02 / 50.0	0.90
Ours (arc, mlp, emo)	0.03 / 50.0	**0.00** / 50.0	0.03 / 50.0	0.02 / 50.0	0.90
Ours (arc, mlp, eye)	**0.00** / 50.0	**0.00** / 50.0	0.03 / 50.0	**0.01 / 50.0**	0.90
Ours (arc, mlp, mix)	**0.00** / 50.0	**0.00** / 50.0	0.03 / 50.0	**0.01 / 50.0**	0.90
Ours (ada, mlp, eye)	3.73 / 70.8	0.43 / 65.5	NA (*c.f.* train/ eval overlap)	2.08 / 68.2	0.87
Ours (mix, mlp, eye)	0.07 / 50.0	0.00 / 50.0		0.04 / 50.0	**0.91**

Impact of ID extraction models. Existing face swapping solutions [93] also leverage out-of-the-box identification networks (*e.g.*, ArcFace [122] as the most common choice), but they do not provide any analysis on the possible bias that these pretrained methods may have and how such bias may impact the de-identification process, *e.g.*, by improperly disentangling some facial features.

To address this concern, we present our analyses in Tables 6.6, 6.7, and 6.8. Our method can harness multiple ID extractors, thus we compare distinct versions of our solutions: employing ArcFace [122], AdaFace [251], SphereFace [306], or a fusion of these methods. Notably, we exclude

Table 6.7: Utility performance comparison of different versions of our methods over diverse downstream tasks on LFW dataset [203].

Methods	Facial landmarks (L2 pixel distance ↓)								Gaze estimation (MAE ° ↓)				Emotion	
	RetinaFace (5 points)				Dlib (68 points)				L2CS Net		ETH XGaze		(Accuracy % ↑)	
	All	Eyes	Nose	Mouth	All	Eyes	Nose	Mouth	Pitch	Yaw	Pitch	Yaw	DAN	DF
Ours (arc, opp, Ø)	**12.5**	**7.6**	**5.4**	**8.0**	**177.6**	**25.8**	**16.1**	**49.8**	7.0	9.2	6.1	12.1	51.2	50.5
Ours (arc, ved, emo)	12.7	**7.6**	5.7	8.1	187.6	26.5	18.8	55.6	7.4	10.8	6.1	13.1	58.8	49.8
Ours (arc, ved, eye)	12.9	7.7	5.6	8.2	203.3	28.8	19.8	60.0	6.8	8.4	5.6	10.0	46.2	47.0
Ours (arc, mlp, emo)	17.1	10.3	7.6	10.8	210.9	31.1	20.3	62.3	7.4	11.3	6.4	14.2	**59.5**	**51.0**
Ours (arc, mlp, eye)	17.3	10.4	7.6	11.2	218.1	31.2	20.1	63.1	7.0	8.0	5.9	9.9	42.9	46.9
Ours (arc, mlp, mix)	16.5	9.9	7.3	10.6	202.4	28.3	18.5	61.4	6.9	7.8	5.5	9.6	**59.5**	49.2
Ours (arc, mlp, eye)	17.3	10.4	7.6	11.2	218.1	31.2	20.1	63.1	7.0	8.0	5.9	9.9	42.9	46.9
Ours (ada, mlp, eye)	15.9	9.3	7.3	10.1	211.4	30.2	21.9	64.9	6.3	**7.3**	**5.4**	9.2	43.8	46.8
Ours (sph, mlp, eye)	14.9	8.9	6.5	9.5	205.2	28.7	21.2	60.5	**6.1**	7.5	**5.4**	10.3	47.4	49.4
Ours (arc+ada, mlp, eye)	15.7	9.3	6.9	10.3	202.8	27.0	18.7	62.1	7.2	7.8	6.2	9.7	42.8	46.9
Ours (arc+sph, mlp, eye)	19.7	11.7	8.4	12.8	230.1	34.0	22.7	66.0	6.8	7.4	5.8	**9.0**	39.1	45.9

the assessment on one ID extractor when it is utilized in the de-identification pipeline, *e.g.* AdaFace in Table 6.6 last tow rows and SphereFace in Table 6.8 last tow rows, ensuring fairness.

Table 6.8 underscores that combining various ID extractors yields enhanced de-identification and non-invertibility. Particularly with AdaFace-based pipelines, this effect is evident. When solely used, AdaFace exhibits slight bias or performance limitations (vis-à-vis FaceNet [413] or SphereFace [306] for re-identification), possibly due to missing biometric features, leading to higher re-identification rates post-obfuscation compared to other ID extractors. However, coupling AdaFace with an alternative ID method like ArcFace [122] mitigates the re-identification rate effectively. Moreover, combining multiple identity extractors notably boosts resilience against inversion attacks (as evident in the last two columns of Table 6.8), as anticipated from mixture-of-experts approaches.

Table 6.8: Re-identification performance and invertibility of different proposed ID transformation methods. (mix1 indicates arc+ada, mix2 indicates arc+sph.)

Methods	TPR (%) @ FPR=1e-3 ↓			
	Swapped		Inverted	
	FaceNet	SphereFace	FaceNet	SphereFace
Ours (arc, mlp, eye)	**0.00**	**0.00**	52.90	45.97
Ours (ada, mlp, eye)	3.73	0.43	49.77	42.63
Ours (mix1, mlp, eye)	0.07	**0.00**	36.70	**34.70**
Ours (sph, mlp, eye)	**0.00**	NA	68.70	NA
Ours (mix2, mlp, eye)	**0.00**		**31.93**	

Nonetheless, a trade-off between preserving privacy and utility remains observable. As Table 6.7 illustrates, solutions leveraging multiple ID extractors tend to exert a slightly greater impact on utility attributes, resulting in a minor accuracy dip for the designated downstream tasks. Navigating this trade-off and devising a solution that better disentangles identity and utility attributes—given their non-overlapping nature—remains an open challenge. Nevertheless, we believe that *Disguise* represents a substantial stride forward in this regard (as evident from comparisons to state-of-the-art in both the main paper and this document).

Impact of ID transformation models. Figure 6.7 extends the analysis presented in Tables 6.4 and 6.5, highlighting the superiority of our VED-based obfuscation scheme compared to the other MLP-based proposed solution, as well as their superiority compared to prior art. The first row in Figure 6.7 shows that compared to other methods, our VED-based model is further from the positive pairs both on the histogram and ROC curve, demonstrating the best de-identification ability. The second

Table 6.9: Evaluation of ID transformation based on noise application only, in terms of de-identification and non-invertibility of the resulting images.

Methods		TPR (%) @ FPR=1e-3 ↓			
		Swapped		Inverted	
Net.	Noise	FaceNet	SphereFace	FaceNet	SphereFace
Ø	$\beta = 0.25$	19.27	12.87	**1.27**	**0.30**
	$\beta = 0.5$	15.37	9.90	1.37	**0.30**
	$\beta = 1.0$	13.03	7.73	1.47	0.43
	$\beta = 2.0$	12.63	6.23	1.93	0.57
	$\beta = 4.0$	**11.30**	6.50	1.47	0.47
	$\beta = 8.0$	**11.30**	**6.27**	2.23	0.50

row shows that when we introduce more β noise in our MLP model, both the histogram and ROC curve move closer towards the positive pairs. When $\beta = 0.5$, our MLP model can de-identify facial images on which recognition model has performance close to random guess. For our VED model, when increasing the α noise, the histogram stays close to the negative pairs and the ROC curve stays close to the diagonal line, as shown in the third row. These results suggest that our VED model achieves the best de-identification while ensuring non re-identifiability.

As a reminder, we define α and β as inversely proportional to ϵ, *c.f.* $\alpha = \frac{\Delta\psi}{\epsilon}$ and $\beta = \frac{\Delta\psi_{\text{mlp}}}{\epsilon}$. As a measure of privacy budget, the higher ϵ is fixed (*i.e.*, the lower α or β), the higher the privacy loss, *c.f.* $\log \mathrm{P}(\widetilde{z}|z) - \log \mathrm{P}(\widetilde{z}|z') \leq \epsilon$ according to the formal definition in Problem Formulation subsection. Local differential privacy (LDP) guarantees that an adversary observing $\widetilde{z}$ cannot determine with some degree of confidence if it comes from z or z'. *E.g.*, $\epsilon = 0$ would mean

zero confidence in linking a masked ID to a specific input one, as only noise would be transferred (*c.f*. Laplacian noise with $\alpha = \frac{\Delta\psi}{\epsilon} = \infty$). To choose ϵ (and thus α) adequately based on privacy budget, one should first estimate the sensitivity Δ of the processing function. Following standard practice [303, 494], we measure the sensitivity of ours empirically: *e.g*., over LFW dataset, we obtain $\Delta\psi_{\text{ved}} = \sup_{z,z'} \|\psi(z) - \psi(z')\|_1 = 33.92$ (*e.g*., hence fixing $\alpha = 2$ means opting for a relative privacy budget equals to $\epsilon = 67.84$).

We enhance the analysis presented in Table 6.5 with additional insights from Table 6.9. This new table illustrates the performance of the ID transformation scheme, which entails applying solely ϵ-controlled Laplace noise to the features without employing additional neural networks for further vector obfuscation. Comparatively, our proposed Multi-Layer Perceptron (MLP) and Variational Encoder-Decoder (VED) solutions distinctly elevate identity obfuscation beyond the effects of noise-only feature manipulation, as depicted in Table 6.9. However, their continuous nature renders them more susceptible to re-identification risk, especially for similar privacy budgets. In cases of slight noise values, they could inadvertently map distinct inputs to a common fabricated identity. Despite this, we maintain the conviction that our VED-based approach strikes the most optimal balance between maximal de-identification and non-reidentifiability.

Comparison with other noise-based ID tampering methods. Some other methods [290, 303, 289, 494], have been recently proposed to tackle de-identification of facial images by extracting identity features from the target data, altering it, and decoding it back into a similar but obfuscated image. While we could not satisfyingly reproduce their results (no implementation has been released), we could approximate their solution using our own framework. Indeed, most of these methods can be described as a subset of our modular solution, *i.e*., minus our main contributions. This is especially

true for DP-Image [303] (not peer-reviewed yet) and IdentityDP [494] (published in August, the 28$^{\text{th}}$, 2022), which use an image encoder-decoder combined with an ID extractor [122] and ID/image feature mixer, similar to ours. However, they do not provide our additional guarantees in terms of disentanglement of the facial attributes and preservation of the utility ones by using mixture-of-experts supervision. More importantly, they obfuscate the extracted ID vector (before injecting it with the residual image features and decoding it back into an image) only by adding Laplace noise to them. They do not leverage additional transformations in the ID latent space to ensure optimal de-identification, such as our MLP and VED neural functions.

To highlight the impact of our proposed ID transformation functions and indirectly compare to these other solutions, we direct the readers to Figure 6.8. For each original image, we display the results obtained by transforming the extracted ID features either after only adding Laplace-based noise to them (first row); after applying our proposed $\psi^{\epsilon}_{\text{mlp}}$, *i.e.*, adding Laplace noise and then passing the vector to our MLP optimized to ensure de-identification (second row); or after passing the vector to our VED $\psi^{\epsilon}_{\text{ved}}$, which also applies ϵ-controlled noise to the data in its own latent space (third row). For each solution, we provide several results with different privacy budgets (β parameter, encompassing ϵ).

We observe that applying only ϵ-controlled noise to the ID vector results in images barely obfuscated (*e.g.*, same nose/cheek/eyebrow shapes) compared to additionally using our proposed neural functions, for the same privacy budgets β. Furthermore, our VED-based solution provides better continuity in the obfuscated results w.r.t. β compared to the other two variants. Such continuity makes choosing an adequate privacy budget much more intuitive and straightforward for users.

Impact of utility experts. Our observations indicate that engaging in fine-tuning alongside utility experts yields notable enhancements in preserving performance for downstream tasks, as evidenced by the findings in Tables 6.6 and 6.7. To delve into specifics, we note that in the gaze estimation task, the model fine-tuned with the inclusion of eye-related utility experts showcases the most minimal offset, denoting superior alignment. Similarly, the same pattern emerges in the context of emotion recognition, where the model fine-tuned with emotion-centric utility experts achieves optimal results. On the other hand, concerning facial landmarks, intriguing dynamics come to light. The model deprived of fine-tuning with emotion or gaze experts demonstrates the highest performance in this regard. This contrast highlights the existence of a trade-off phenomenon, indicating that distinct utility experts exert varying degrees of influence, necessitating a balanced consideration.

6.5 Conclusion and Discussion

We introduced *Disguise*, a privacy-enhancing face de-identification model that ensures both depicted people's privacy and image usability. Our experiments demonstrate its effectiveness in pre-processing sensitive data for inference or training. Rooted in privacy and mixture-of-experts theory, it outperforms prior methods in re-identification robustness and utility preservation.

Limitations. Note that our model is tailored for face obfuscation and does not address other identity-revealing visual attributes (*e.g.*, distinctive glasses, haircuts, backgrounds). Broader ID-extracting methods like $H_{\mathcal{Z}}$ [44] could potentially handle this. Additionally, *Disguise* might benefit from multi-objective learning research [124, 344] to optimize cases where identity and utility features overlap.

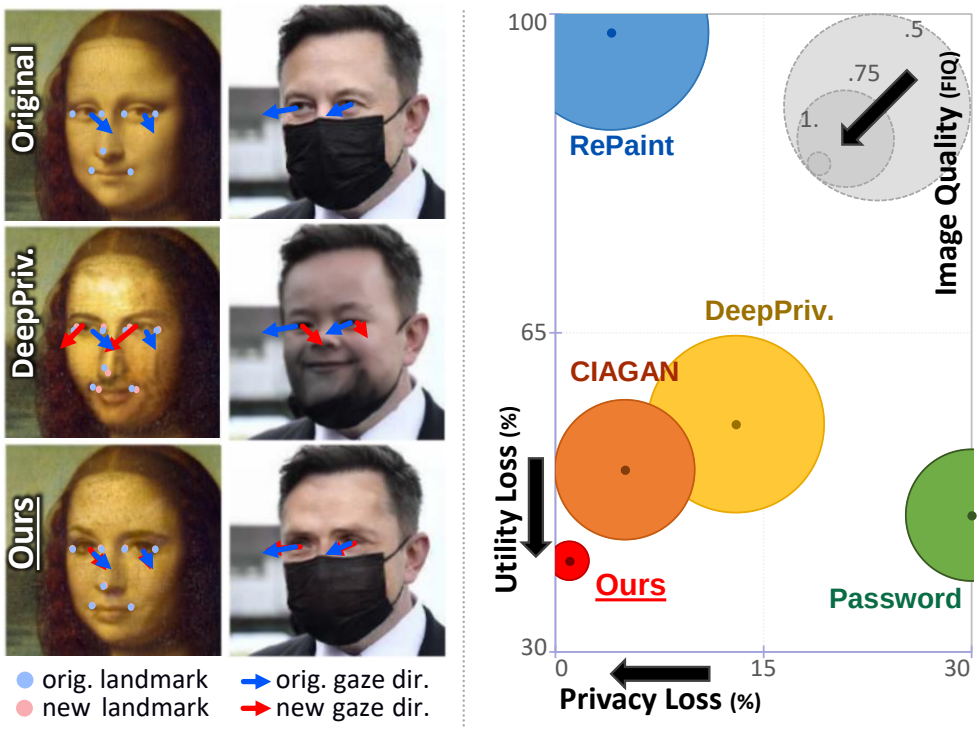


Figure 6.1: *Disguise* anonymizes face images while preserving their utility (i.e., attributes relevant to downstream tasks). For instance, facial landmarks and gaze direction are better preserved compared to existing methods, as shown in the figure that the red dots for landmarks and red arrows for gazes in the new images are more aligned with the blue ones in the original images. We outperform prior art by a large margin along various axes, including privacy, utility, and image quality. For image quality, small radius indicates higher FIQ [448] score and better image quality.

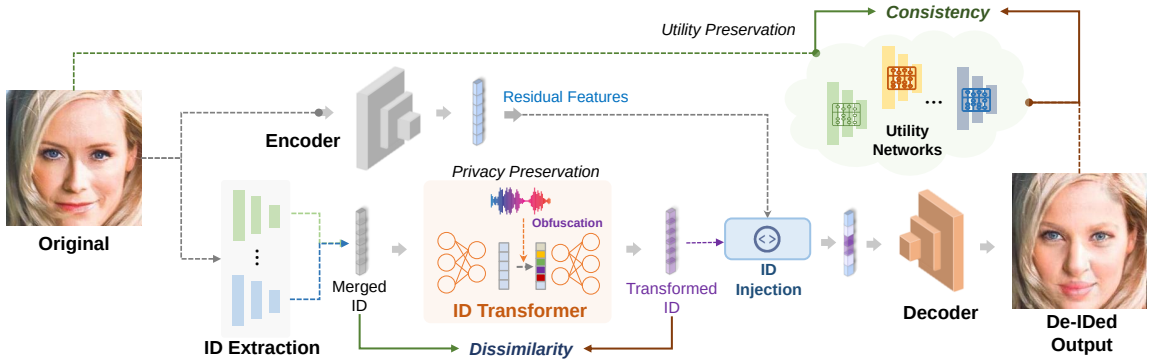


Figure 6.2: Illustration of the training process for the proposed *Disguise* framework. More discussions in Methodology Section.

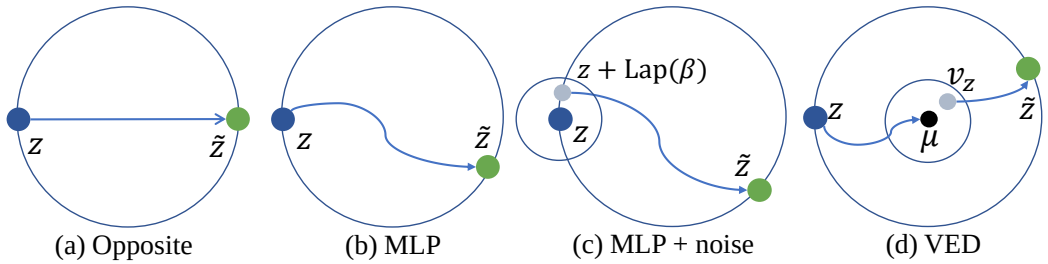


Figure 6.3: Identity transformation. The identity vector is normalized to the surface of a unit n-sphere.

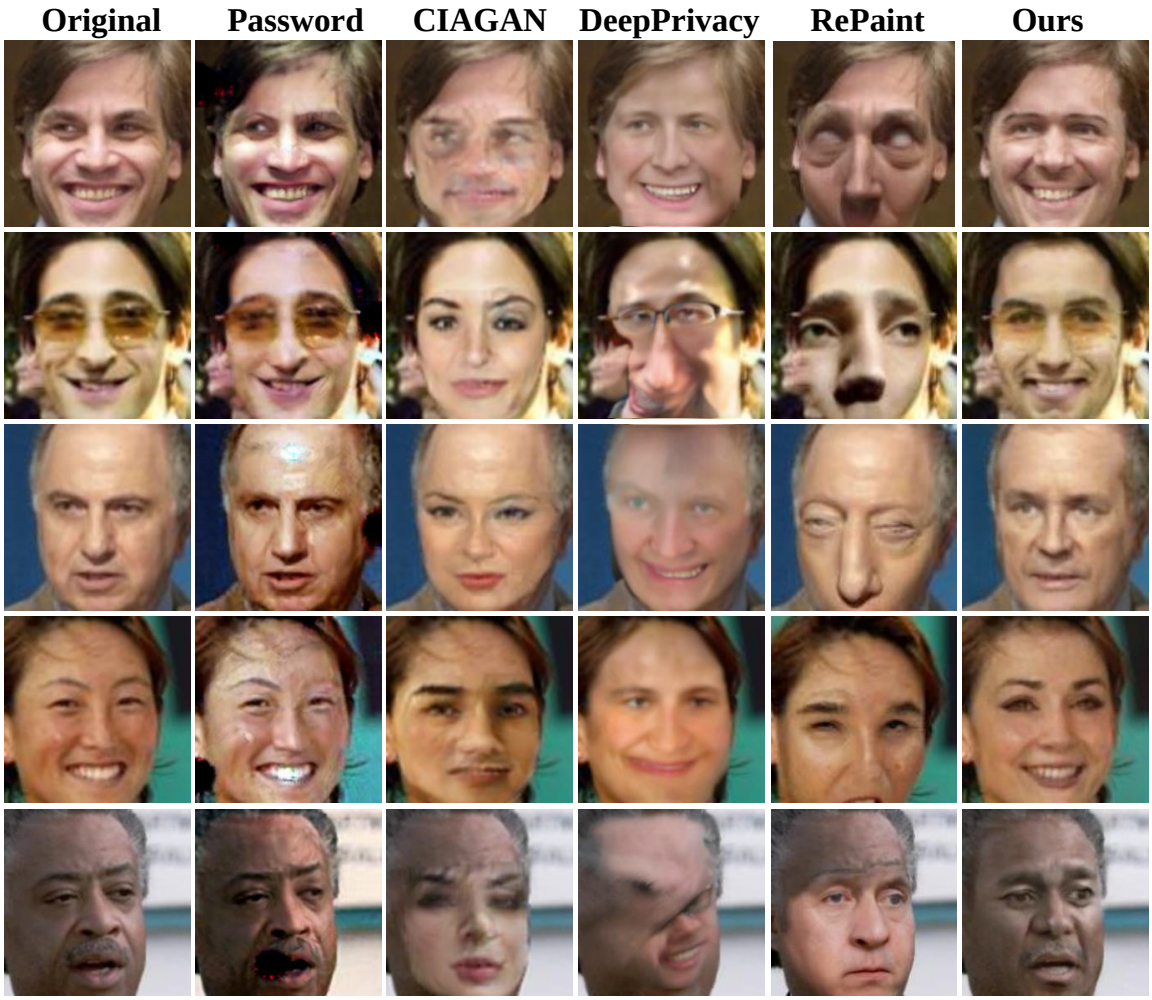


Figure 6.4: Qualitative results of different methods. Ours preserves utility while anonymizing identities.

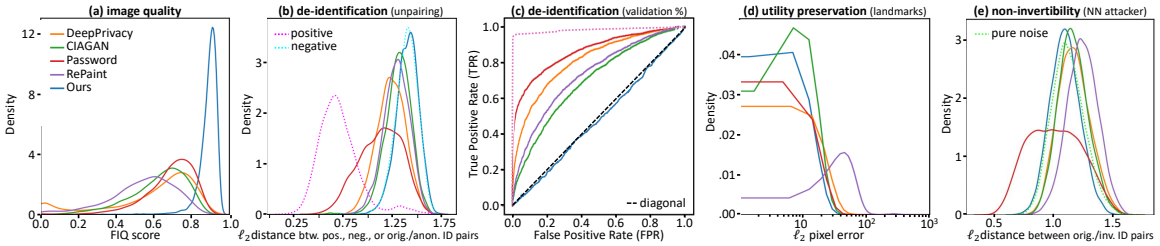


Figure 6.5: *Disguise* outperforms existing methods in various aspects, including image quality, de-id rate, and utility. For non-invertibility, our solution is close to other methods that completely erase the original IDs (*i.e.*, recovering pure Gaussian noise).

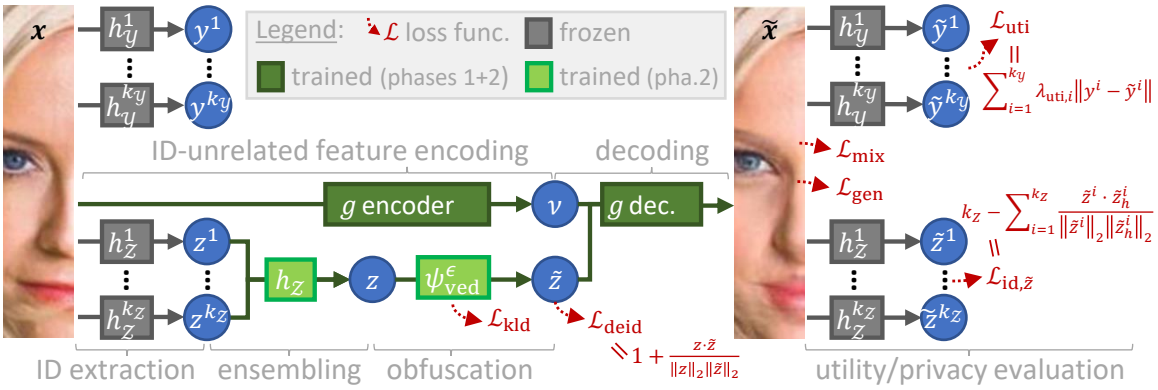


Figure 6.6: Detailed training pipeline of *Disguise*, in supplement to Figure 6.2. The proposed solution is end-to-end differentiable. However, in practice, to guide the optimization process, we train the network in two phases. Firstly, we train the face-swapping network (the branch marked in dark green); then in the second phase, we add the ID obfuscation branch (marked in light green) and the utility-guaranteeing module (the branch on top) to finetune the whole network.

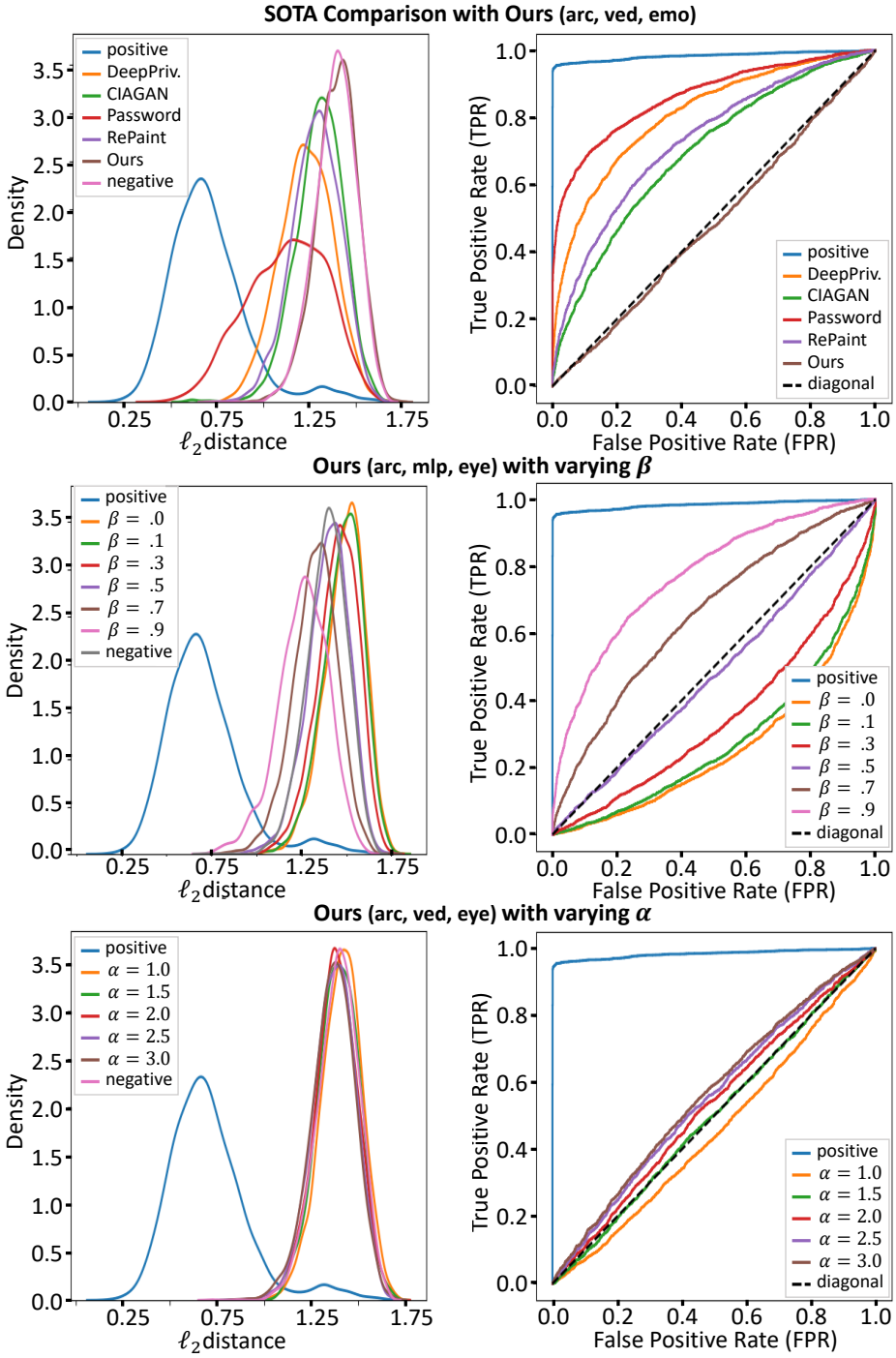


Figure 6.7: Left: Histogram of ℓ_2 distances between positive, negative, and original-anonymized pairs from LFW set. Right: ROC curves of validation rate for images altered by various methods.

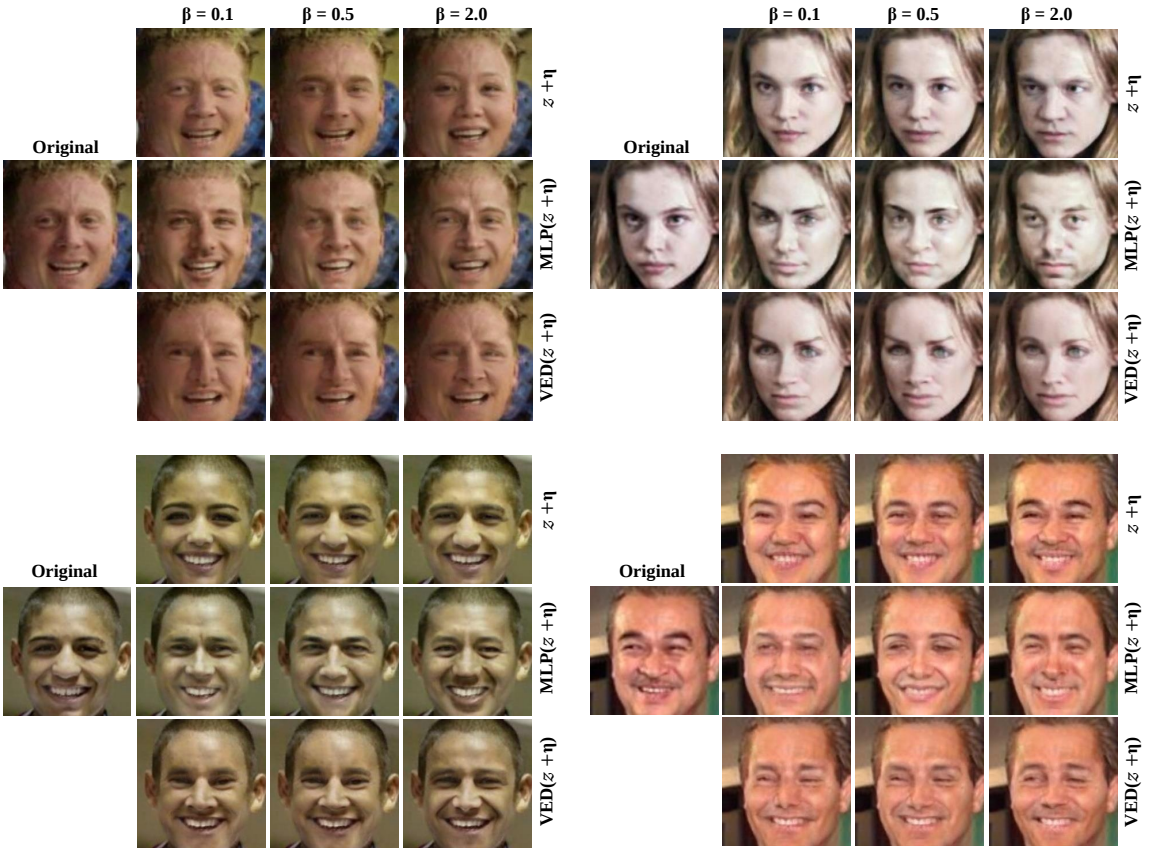


Figure 6.8: Comparison of different ID transformation functions in terms of their impact on the resulting obfuscated images. We compare (1) applying only Laplace noise to the extracted ID vectors (noted "$z+\eta$" in the figure), (2) applying our proposed $\psi^{\epsilon}_{\text{mlp}}$, *i.e.*, applying noise and our MLP (noted "$\text{MLP}(z+\eta)$"), or (3) applying our $\psi^{\epsilon}_{\text{ved}}$, *i.e.*, applying noise and our VED (noted "$\text{VED}(z+\eta)$" here).

Chapter 7

Solving Phase Retrieval with a Learned Reference

7.1 Introduction

The problem of *phase retrieval* refers to the challenge of recovering a real- or complex-valued signal from its amplitude measurements. This problem arises in diffraction imaging, X-ray crystallography, and ptychography [150, 157, 182, 341, 424]. Fourier phase retrieval is a special class of phase retrieval problems aimed at the recovery of a signal from the amplitude of its Fourier coefficients. Let us assume that Fourier amplitude measurements are given as

$$y = |Fx| + \eta, \tag{7.1}$$

where F denotes the Fourier transform operator, x denotes the unknown signal or image, and η denotes the measurement noise. Our goal is to recover x given y.

Fourier phase retrieval is essential in many applications, especially in optical coherent imaging. Classical methods for phase retrieval utilize the prior knowledge about the support and positivity of the signals [150, 157]. Subsequent work has considered the case where the unknown signal is *structured* and belongs to a low-dimensional manifold that is known *a priori*. Examples of such low-dimensional structures include sparsity [476, 230], low-rank [229, 99], or neural generative models [228, 231]. Other techniques like Amplitude flow [477] and Wirtinger flow use alternating minimization [68]. Many of these newer algorithms involve solving a *non-convex* problem using iterative, gradient-based methods; therefore, they need to be carefully initialized. The initialization technique of choice is spectral initialization, first proposed in the context of phase retrieval in [360], and extended to the sparse signal case in [476, 230].

Fourier phase retrieval problem does not satisfy the assumptions needed for successful spectral initialization and remains highly sensitive to the initialization choice. Furthermore, Fourier amplitude measurements have the so-called trivial ambiguities about possible shifts and flips of the images. Therefore, many Fourier phase retrieval methods test a number of random initializations with all possible flips and shifts and select the estimate with the best recovery error [339].

In this paper, we assume that a known (learned) reference is added to the signal before capturing the Fourier amplitude measurements. The main motivation for this comes from the empirical observation that knowing a part of the image can often help resolve the trivial ambiguities [32, 176, 214]. We extend this concept and assume that a known reference signal is added to the target signal and aim to recover the target signal from the Fourier amplitude of the combined signal. Adding a reference may not feasible in all cases, but our method will be applicable whenever we can add a reference or split the target signal into known and unknown parts. We can describe the

Fourier amplitude (phaseless) measurements with a known reference signal u as

$$y = |F(x + u)| + \eta. \tag{7.2}$$

Similar reference-based measurements and phase retrieval problems also arise in holographic optical coherence imaging [364].

Our goal is to recover the signal x from the amplitude measurements in (7.2). To do that, we implement a gradient descent method for phase retrieval. We present the algorithm as an unrolled network for a general system in Fig. 7.1. Every layer of the network implements one step of the gradient descent update. To minimize the computational complexity of the recovery algorithm, we seek to minimize the number of iterations (hence the layers in the network). In addition, we seek to learn the reference u to maximize the accuracy of the recovered signal for a given number of iterations. The learned u and reconstruction results for different datasets are summarized in Fig. 7.2.

7.1.1 Our Contributions

We present an iterative method to efficiently recover a signal from the Fourier amplitude measurements using a fixed number of iterations. To achieve this goal, we first learn a reference signal that can be added to the phaseless Fourier measurements to enable the exact solution of the phase retrieval problem. We demonstrate that the reference learned on a very small training set perform remarkably well on the test dataset.

Our main contributions can be summarized as follows.

- The proposed method uses a fixed number of gradient descent iterations (i.e., fixed computational cost) to solve the Fourier phase retrieval problem.

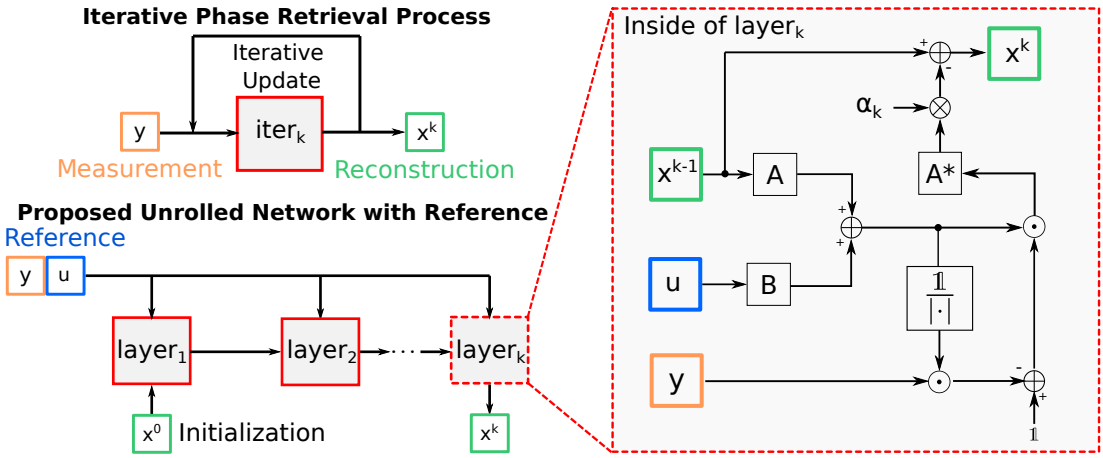


Figure 7.1: Our proposed approach for learning reference signal by solving phase retrieval using an unrolled network. Unrolled network has K layers. Each layer$_k$ gets amplitude measurements y, reference u, and estimate x^{k-1} as inputs, and updates the estimate to x^k. The operations inside layer$_k$ are shown in the dashed box on the right, where A and B are both linear measurement operators, and A^* is the adjoint operator of A.

- We formulate the gradient descent method as an unrolled network that allows us to learn a robust reference signal for a class of images. We demonstrate that reference learned on a very small dataset performs remarkably well on diverse and large test datasets. To the best of our knowledge, this is the first work on learning a reference for phase retrieval problems.

- We tested our method extensively on different challenging datasets and demonstrated the superiority of our method.

- We demonstrate the robustness of our approach by testing it with the noisy measurements using the reference that was trained on noise-free measurements.

7.2 Related Work

Holography. Digital holography is an interferometric imaging technique that does not require the use of any imaging lens. Utilizing the theory of diffraction of light, a hologram can be used to reconstruct three-dimensional (3D) images [376]. With this advantage, holography can be used to perform simultaneous imaging of multidimensional information, such as 3D structure, dynamics, quantitative phase, multiple wavelengths, and polarization state of light [440]. In the computational imaging community, many attempts have been made in solving holographic phase retrieval using references, among which [32] has been very successful. Motivated by the reference design for holographic phase retrieval, we are trying to explore a way to design references for general phase retrieval.

Phase Retrieval. The phase retrieval problem has drawn considerable attention over the years, as many optical detection devices can only measure amplitudes of the Fourier transform of the underlying object (signal or image). Fourier phase retrieval is a particular instance of this problem that arises in optical coherent imaging, where we seek to recover an image from its Fourier modulus [150, 157, 400, 341, 424, 327]. Existing algorithms for solving phase retrieval can be broadly classified into convex and non-convex approaches [215]. Convex approaches usually solve a constrained optimization problem after lifting the problem. The PhaseLift algorithm [69] and its variations [167], [67] belong to this class. On the other hand, non-convex approaches usually depend on Amplitude flow [476, 475] and Wirtinger flow [68, 531, 98, 59]. If we know some structure of the signal a priori, it helps in the reconstruction. Sparsity is a very popular signal prior. Some of the approaches for sparse phase retrieval include [366, 292, 23, 225, 360, 59, 476]. Furthermore, [360, 230, 215] used minimization (AltMin)-based approach and [85] used total variation regu-

larization to solve phase retrieval. Recently, various researchers have explored the idea of replacing the sparsity priors with generative priors for solving inverse problems. Some of the generative prior-based approaches can be found in [215, 231, 180, 421].

Data-Driven Approaches for Phase Retrieval. The use of deep learning-based methods to solve computational imaging problems such as phase retrieval is becoming popular. Deep learning methods leverage the power of huge amounts of data and tend to provide superior performance compared to traditional methods while also run significantly faster with the acceleration of GPU devices. A few examples demonstrating the benefit of the data-driven approaches include [339] for robust phase retrieval, [246] for Fourier ptychographic microscopy, and [399] for holographic image reconstruction.

Unrolled Network for Inverse Problem. Unrolled networks, which are constructed by unrolled iterations of a generic non-linear reconstruction algorithm, have also been gaining popularity for solving inverse problems in recent years [247, 127, 166, 481, 179, 520, 240, 53]. Iterative methods usually terminate the iteration when the condition satisfies theoretical convergence properties, thus rendering the number of iterations uncertain. An unrolled network has a fixed number of iterations (and cost) by construction and they produce good results in a small number of steps while enabling efficient usage of training data.

Reference Design. Fourier phase retrieval faces different trivial ambiguities because of the structure of Fourier transformation. As a phase shift in the Fourier domain results in a circular shift in the spatial domain, we will get the same Fourier amplitude measurements for any circular shift of the original signal. In recent papers [32, 527, 176, 214], authors tried to use side information with sparsity prior to mitigate these ambiguities. However, in those studies, the reference and target

signal are separated by some margin. If the separation between target and reference is large enough, then the nonlinear PR problem simplifies to a linear inverse problem [14, 32].

In this paper, we consider the reference signal to be additive and overlapping with the target signal. To the best of our knowledge, there has not been any study on such unrestricted reference design. While driven by data, our approach for reference design uses training samples in a very efficient way. The number of training images required by our network is parsimonious without limiting its generalizability. The reference learned by our network provides robust recovery test images with different sizes. Apart from the great flexibility, our unrolled network uses a well-defined routine in each layer and demonstrates excellent interpretability as opposed to black-box deep neural networks.

7.3 Proposed Approach

We use the general formulation for the phase retrieval from amplitude measurements. The formulation can be extended for phase retrieval with squared amplitude measurement as well. In our setup, we model amplitude measurements of a target signal x and a reference signal u as $y = |Ax + Bu|$, where A and B are linear measurement operators. Our goal is to learn a reference signal that provides us the best recovery of the target signal. We formulate this overall task as the following optimization problem:

$$\underset{\hat{x}(u)}{\text{minimize}} \ \|x - \hat{x}(u)\|_2^2 \quad \text{s.t.} \ \ y = |A\hat{x}(u) + Bu|, \tag{7.3}$$

where $\hat{x}(u)$ denotes the solution of the phase retrieval problem for a given reference u. Our approach to learn u and solve (7.3) can be divided into two nested steps: (1) Outer step updates u to minimize

the recovery error for phase retrieval and (2) inner step uses the learned u to recover target images by solving phase retrieval.

To solve the (inner step) of phase retrieval problem, we use an unrolled network. Figure 7.1 depicts the structure of our phase retrieval algorithm. In the unrolled phase retrieval network, we have K blocks to represent K iterations of the phase retrieval algorithm. We minimize the following loss to solve the phase retrieval problem:

$$L_x(x, u) = \|y - |Ax + Bu|\|_2^2. \tag{7.4}$$

Every block of the unrolled phase retrieval network is equivalent to one gradient descent step for (7.4). For some value of reference estimate, u, we can represent the target signal estimate after $k + 1^{th}$ block of the unrolled network as

$$x^{k+1} = x^k - \alpha_k \nabla_x L_x(x^k, u), \tag{7.5}$$

where $\nabla_x L_x(x^k, u)$ is the gradient of L_x with respect to x at the given values of x^k, u. As the loss function in (7.4) is not differentiable, we can redefine it as

$$L_x(x, u) = \|y \odot p - (Ax + Bu)\|_2^2, \tag{7.6}$$

where $p = \angle(Ax^k + Bu) = (Ax^k + Bu)/|Ax^k + Bu|$. The expression of gradient can be written as

$$\nabla_x L_x(x^k, u) = 2A^*[p \odot (p^* \odot (Ax^k + Bu) - y)], \tag{7.7}$$

where A^* denotes the adjoint of A. After K blocks, we get the estimate of the target signal that we denote as $\hat{x}(u) = x^K$.

In the learning phase, we are given a set of training signals, $\{x_1, x_2, ..., x_N\}$, which share the same distribution as our target signals. We initialize x^0 and u^0 with some initial (feasible) values.

First we minimize the following loss with respect to u:

$$L_u(u) = \sum_{i=1}^{N} \|x_i - \hat{x}_i\|_2^2 = \sum_{i=1}^{N} \|x_i - x_i^K\|_2^2. \tag{7.8}$$

We can rewrite (7.8) using the gradient recursion in (7.5) as

$$L_u(u) = \sum_{i=1}^{N} \|x_i - x_i^0 + \sum_{k=0}^{K-1} \alpha_k \nabla_x L_x(x_i^k, u)\|_2^2. \tag{7.9}$$

We can then use gradient descent to to minimize $L_u(u)$. We can represent the $j+1^{th}$ iteration of gradient descent step as

$$u^{j+1} = u^j - \beta \nabla_u L_u(u^j). \tag{7.10}$$

The expression for $\nabla_u L_u(u)$ can be written as

$$\nabla_u L_u(u) = 2 \sum_{i=1}^{N} \left[\sum_{k=0}^{K-1} \alpha_k J_u(x_i^k, u) \right] \left[x_i - x_i^0 + \sum_{k=0}^{K-1} \alpha_k \nabla_x L_x(x_i^k, u) \right], \tag{7.11}$$

where $J_u(x_i^k, u) = \nabla_u \nabla_x L_x(x_i^k, u)$ is a Jacobian matrix with rows and columns of the same size as u and x, respectively. The measurement vector $y = |Ax + Bu|$ is a function of u during training. Since we model $\hat{x}(u)$ as an unrolled network, we can think of the gradient step as a backpropagation step. To compute $\nabla_u L_u(u)$, we backpropagate through the entire unrolled network. At the end of J^{th} outer iteration, we will get our learned reference $\hat{u} = u^J$.

Once we have learned a reference, $\hat{u}$, we can use it to capture (phaseless) amplitude measurements as $y = |Ax^* + B\hat{u}|$ for target signal x^*. To solve the phase retrieval problem, we perform one forward pass through the unrolled network. Pseudocodes for training and testing are provided in Algorithms 7,6.

In our Fourier phase retrieval experiments $A = B = F$, where F is the Fourier transform operation. To implement similar method for squared amplitude measurements, we can simply

Algorithm 5 Learning Reference Signal

Input: Training signals $\{x_1, x_2, ..., x_N\}$, measurement operators A and B

Output: Optimal reference $\hat{u} = u^J$

Initialize $\{x_1^0, x_2^0, ..., x_N^0\}, u^0$

for $j = 0, 1, ..., J-1$ **do**

  for $i = 1, 2, ..., N$ **do**

    $y_i = |Ax_i^* + Bu^j|$

    for $k = 0, 1, ..., K-1$ **do**

      $L_x(x_i^k, u^j) = \|y_i - |Ax_i^k + Bu^j|\|_2^2$

      $x_i^{k+1} \leftarrow x_i^k - \alpha_k \nabla_x L_x(x_i^k, u^j)$

    end for

  end for

  $L_u(u^j) = \sum_{i=1}^N \|x_i^* - x_i^0 + \sum_{k=1}^K \alpha_k \nabla_x L_x(x_i^{k-1}, u^j)\|_2^2$

  $u^{j+1} \leftarrow u^j - \beta \nabla_u L_u(u^j)$

end for

return $\hat{u} = u^J$

replace $p = \angle(Ax^k + Bu^j)$ with $p = Ax^k + Bu^j$. In all our experiments, we initialized x^0 as a zero vector whenever $\hat{u} \neq 0$. We can also add additional constraints on the reference while minimizing the loss function in (7.9). In our experiments, we used target signals with intensity values in the range $[0, 1]$; therefore, we restricted the range of entries in u to $[0, 1]$ as well. We discuss other constraints in the experiment section.

Algorithm 6 Solving Phase Retrieval via Unrolled Network

Input: Measurements y, learned reference $\hat{u}$, measurement operators A and B

Output: Estimation of target signal $\hat{x} = x^K$

Initialize x^0

for $k = 0, 1, ..., K-1$ **do**

  $L_x(x^k, \hat{u}) = \|y - |Ax^k + B\hat{u}|\|_2^2$

  $x^{k+1} \leftarrow x^k - \alpha_k \nabla_x L_x(x^k, \hat{u})$

end for

return $\hat{x} = x^K$

7.4 Experiments

Datasets. We have used MNIST digits, EMNIST letters, Fashion MNIST, CIFAR10, SVHN, CelebA datasets, and different well-known standard images for our experiments. We convert all images to grayscale and resize 28×28 images to 32×32. Although there are tens of thousands training images in MNIST, EMNIST letters, Fashion MNIST, CIFAR10, and SVHN dataset, we have used only a few (i.e.. 32) of them in training. We have shown that the references learned on the small number of training images perform remarkably well on the entire test dataset. MNIST, Fashion MNIST, and CIFAR10 test datasets contain 10000 test images each; EMNIST letters dataset contains 24800 test images; SVHN test dataset contains 26032 test images. We used 1032 images from CelebA and center-cropped and resized all of them to 200×200. We selected 32 images for training and the rest for testing.

We present the results for these different datasets using references learned from 32 images from the same dataset in Fig. 7.2. We present results for six standard images of size 512×512 from [339] using a resized reference learned from CelebA dataset in Fig. 7.3.

Measurements. We simulated amplitude measurements of the 2D Fourier transform. We performed 4 times oversampling in the spatial domain for both reference and target signal. Unless otherwise mentioned, we consider our measurements to be noise-free. We also report results for noisy measurements.

7.4.1 Configurations of Reference (u)

The reference signal u, which we are trying to learn, has a number of hyper-parameters that inherently affect the performance of the phase retrieval process. We considered several constraints on u, including the support, size, range, position, and sparsity.

We tested reference signals with both complex and real values and found that u has comparable results in the two domains. Since it is easy to physically create amplitude or phase-only reference signals, we constrain u to be in the real domain; thus, $u \in \mathbb{R}^{m \times n}$ and m, n represent height and width, respectively. The height and width of u determine the overlapping area between the target signal and the reference. We found that u with larger size tends to have better performance, especially when the value of u is constrained to a small range. The intensity values of u play a major role in its performance. If we constrain the value of u to be within a certain range: $u[i, j] \in [u_{min}, u_{max}]$, for all i, j, we observed that bigger range of u yields better performance. This is because when u is unconstrained then we can construct a u with a large norm. Consider the noiseless setting with quadratic measurements $|F(x+u)|^2 = |Fx|^2 + |Fu|^2 + 2\text{Re}(Fx \odot Fu)$, the last term is the real value of the element-wise product of target and reference Fourier transforms.

Table 7.1: PSNR for different training size

TRAIN/TEST	MNIST	EMNIST	F. MNIST	SVHN	CIFAR10
TRAINING SIZE=32	66.54	58.72	57.81	57.51	41.60
TRAINING SIZE=128	76.25	64.16	55.86	59.50	44.34
TRAINING SIZE=512	79.14	62.34	52.01	59.78	48.90

We can remove $|Fu|^2$ because it is known. If u is large compared to x, then we can also ignore the quadratic term $|Fx|^2$ and recover x in a single iteration if all entries of Fu are nonzero. To avoid this situation and make the problem stable in the presence of noise, we restricted the values in the reference u to be in [0,1] range.

7.4.2 Setup of Training Samples and Sample Size

We observed that we can learn the reference signal from a small number of training images. In Table 7.1, we report test results for different reference signals learned on first N images from MNIST training dataset for $N = 32, 128, 512$. We kept the signal and reference strength (i.e., the range of the signal) equal for this experiment. We observe that increasing the training size improves test performance. However, we can get reasonable reconstruction performance on large test datasets (10k+ images) with reference learned using only 32 images.

7.4.3 Generalization of Reference on Different Classes

We are interested in evaluating the generalization of our learned reference. (i.e., how the reference performs when trained on one dataset and tested on another). In the comparison study, we took the reference u trained on each dataset and then tested them on the remaining 4 datasets. The value range of the reference is between $[0, 1]$, the number of steps in the unrolled network is $K = 50$. We observed that when the datasets share great similarity (e.g., MNIST and EMNIST are both sparse digits or letters), the reference signal tends to work well on both datasets. Even when the datasets differ greatly in their distributions, the reference trained on one dataset provides good results on other datasets (with only a few dB of PSNR decrease in performance).

We also tested our method on shifted and rotated versions of test images. Results in Fig. 7.4 demonstrate that even though the reference was trained on upright and centered images, we can perfectly recover shifted and rotated images.

Our key insight about this generalization phenomenon is that the main challenge in Fourier phase retrieval methods is initialization and ambiguities that arise because of symmetries. We are able to solve these issues using a learned reference because of the following reasons: (1) A reference gives us a good initialization for the phase retrieval iterations. (2) The presence of a reference breaks the symmetries that arise in Fourier amplitude measurements. Moreover, we are not learning to solve the phase retrieval problem in an end-to-end manner or learn a signal-dependent denoiser to solve the inverse problem [339, 399]. We are learning reference signals to primarily help a predefined phase retrieval algorithm to recover the true signal from the phaseless measurements. Thus, the references learned on one class of images provide good results on other images, see Table 7.2. This study shows that the reference learned using our network has the ability to generalize

Table 7.2: PSNR of the Same Reference Tested on Different Datasets

TRAIN/TEST	MNIST	EMNIST	F. MNIST	SVHN	CIFAR10
MNIST	66.54	55.12	40.87	41.87	31.72
EMNIST	**72.84**	**58.72**	52.18	55.42	48.16
F. MNIST	40.87	55.67	**57.81**	50.70	42.85
SVHN	41.87	46.76	49.60	**57.51**	**51.54**
CIFAR10	31.72	38.93	36.40	40.36	41.60

to new datasets, thus making our method suitable for real-life applications where new test cases keep emerging.

7.4.4 Noise Response

To test the robustness of our method in the presence of noise, we added Gaussian and Poisson noise at different levels to the measurements. Poisson noise or shot noise is the most common in the practical systems. We model the Poisson noise following the same approach as in [339]. We simulate the measurements as

$$y(i) = |z(i)| + \eta(i) \quad \text{for all } i = 1, 2, \ldots, m, \tag{7.12}$$

where $\eta(i) \sim \mathcal{N}(0, \sigma^2)$ for Gaussian noise and $\eta(i) \sim \mathcal{N}(0, \lambda|z(i)|)$ for Poisson noise with $z = Ax + Bu$. We varied σ, λ to generate noise at different signal-to-noise ratios. Poisson noise affects the larger measurements with higher strength than the smaller measurements. As the sensors can measure only positive measurements, we kept the measurements positive by applying ReLU

function after noise addition. We can observe the effect of noise in Fig. 7.5. Even though we did not add noise during training, we get reasonable reconstruction and performance degrades gracefully with increased noise.

7.4.5 Random Reference versus Learned Reference

To demonstrate the advantage of the learned reference signal, we compared the performance of learned reference and random reference on some standard images. The results are shown in Fig. 7.3. The learned reference is trained using 32 images from CelebA dataset which we resized to 200×200. The test images used in Fig. 7.3 are 512×512, so we resized the learned reference from 200×200 to 512×512. For random reference, we selected the entries of the reference uniformly at random from $[0, 1]$. We selected the best result out of 100 trials for every test image with random reference. We can observe from the results that our learned reference significantly outperforms the random reference even though the test image distribution is distinct from the training data. The number of steps of the unrolled network is $K = 50$.

7.4.6 Comparison with Existing Phase Retrieval Methods

We have shown comparison with other approaches in Table 7.3. We selected Kaczmarz [488] and Amplitude flow [98] for comparison using PhasePack package [84]. We also show Hybrid Input Output (HIO), which is similar to our phase retrieval routine without any reference. We observe that our approach with learned reference can outperform all other approaches on all the datasets. All the traditional phase retrieval methods suffer from the trivial circular shift, rotation, and flip ambiguities, thus produce significantly worse reconstruction than our method does. Our

Table 7.3: Comparison with Existing Phase Retrieval Methods

METHODS	MNIST	EMNIST	F. MNIST	SVHN	CIFAR10
HIO	9.04	8.42	9.65	19.87	14.70
AMPLITUDE FLOW	9.99	9.79	11.90	20.25	15.04
KACZMARZ	11.81	11.47	13.44	19.48	15.01
FLAT REFERENCE	18.21	17.24	16.56	20.89	15.81
RANDOM REFERENCE	36.87	28.41	27.27	36.45	25.57
LEARNED REFERENCE (OURS)	**66.54**	**58.72**	**57.81**	**57.51**	**41.60**

method uses a reference signal to simplify the initialization and removes the shift/reflect ambiguities. To mathematically explain this fact, a shifted or flipped version of x would not give us the same Fourier measurements as $|F(x+u)|$ if u is chosen appropriately as we do with the learning procedure. As we showed in Fig. 7.5, our method can perfectly recover the shifted and flipped versions of the images using the reference that was trained with upright and centered images.

7.4.7 Effects of Number of Layers (K)

We tested our unrolled network with different numbers of layers (i.e., K) at training and test time. The results are summarized in Fig. 7.6. We first used the same values of K for training and testing. We observed that as K increases, the reconstruction quality (measured in PSNR) improves. Then we fixed $K = 1$ or $K = 10$ at training, but used different values of K at testing. We observed that if we increase K at the test time, PSNR improves up to a certain level and then it plateaus. The PSNR achieved with reference trained with $K = 10$ is better than what the referenced trained with

$K = 1$ provided. These results provide us a trade-off between the reconstruction speed and quality. As we increase K, the reconstruction quality improves but the reconstruction requires more steps (computations and time).

Finally, we learned a reference using $K = 1$ and tested it on different images with $K = 1$. To our surprise, our method was able to produce reasonable quality reconstruction with this extreme setting. We present some single-step reconstructions of each data set in Fig. 7.7.

7.4.8 Localizing the Reference

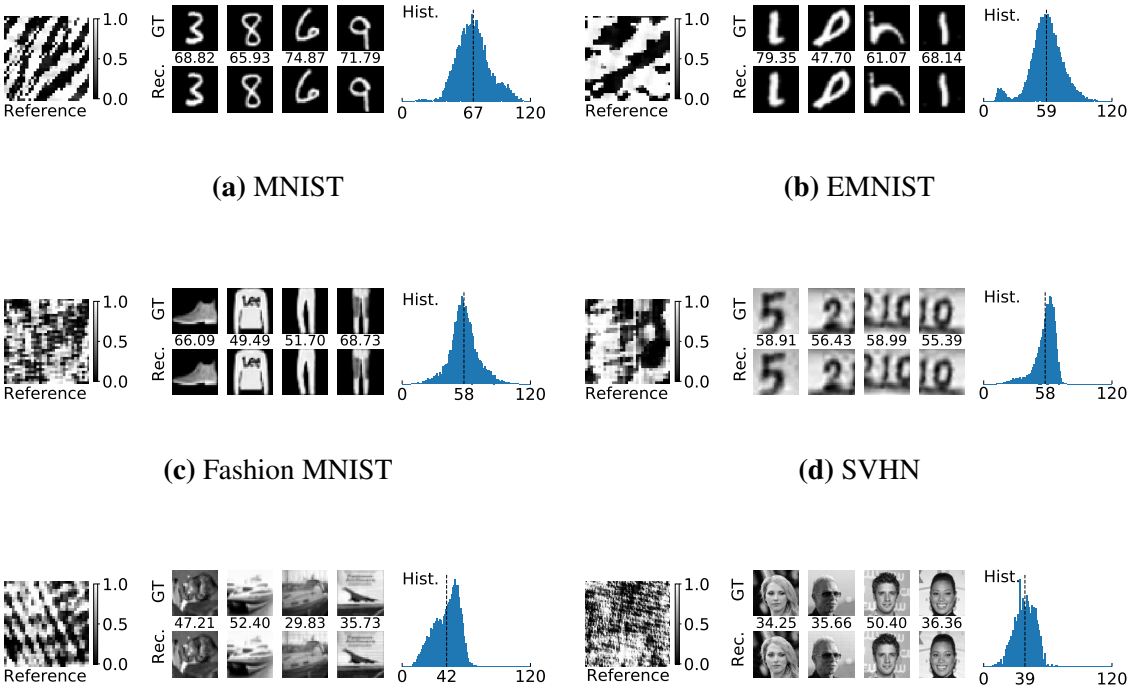
We also evaluated the effect of localizing the reference to a small region. For example, the reference is constrained to be within a small block in the corner or the center of the target signal. We restricted u to be an 8×8 block and placed it in different positions. We found that corner positions provide better results as shown in Fig. 7.8. As we bring the reference support closer to the center, the quality of reconstruction deteriorates. This observation is related to the method in [32, 176, 14], where if the known reference signal is separated from the target signal, then the phase retrieval problem can be solved as a linear inverse problem.

Note that signal recovery from Fourier phase retrieval is equivalent to signal recovery from its autocorrelation. We can write the autocorrelation of target plus reference signals as $(x + u) \star (x + u) = x \star x + u \star u + x \star u + u \star x$. The first term is a quadratic function of x, the second term is known, and the last two terms are linear functions of x. If the supports for x and u are sufficiently separated, then we can separate the last two linear terms from the first two quadratic

terms and recover x by solving a linear problem. However, if x and u have a significant overlap, then we need to solve a nonlinear inverse problem as we do in this paper.

7.5 Conclusion

We presented a framework for learning a reference signal to solve the Fourier phase retrieval problem. The reference signal is learned using a small number of training images using an unrolled network as a solver for the phase retrieval problem. Once learned, the reference signal serves as a prior which significantly improves the efficiency of the signal reconstruction in the phase retrieval process. The learned reference generalizes to a broad class of datasets with different distribution compared to the training samples. We demonstrated the robustness and efficiency of our method through extensive experiments.



(a) MNIST **(b)** EMNIST

(c) Fashion MNIST **(d)** SVHN

(e) CIFAR10 **(f)** CelebA

Figure 7.2: Reconstruction results using learned references. Each block **(a)-(f)** shows results for different dataset: (**left**) learned reference with a colorbar; (**middle**) sample original images and reconstruction with PSNR on top; (**right**) histogram of PSNR over the entire test dataset (vertical dashed line represents the mean PSNR).

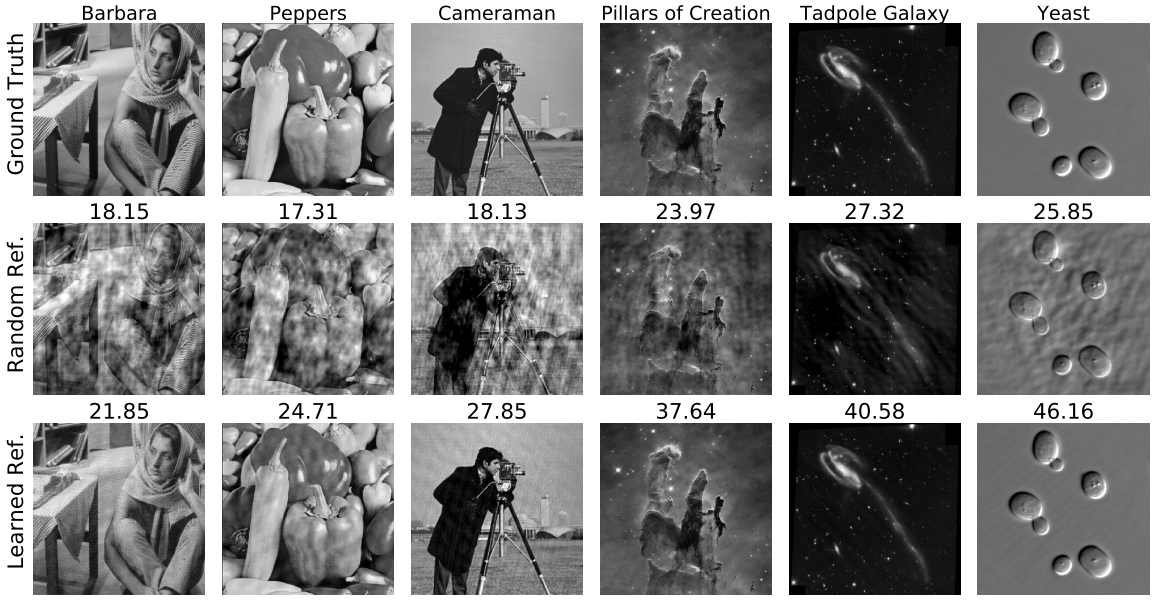
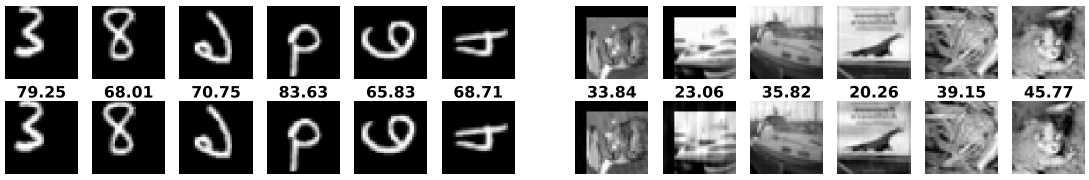
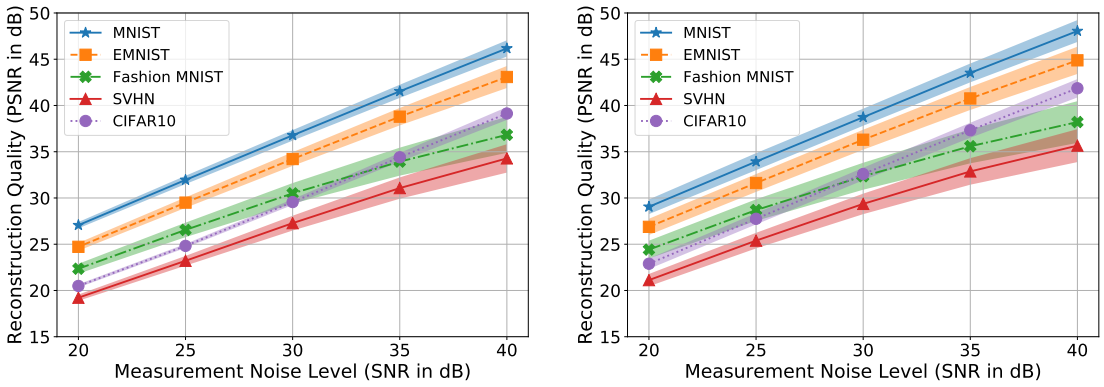


Figure 7.3: Phase retrieval results using learned and random references. **First Row:** Original 512×512 test images. **Second Row:** Reconstruction using random references with uniform distribution between $[0, 1]$ **best result out of 100 trials**. **Third Row:** Reconstruction using the reference learned on CelebA dataset and resized from 200×200 to 512×512. (PSNR shown on top of images).



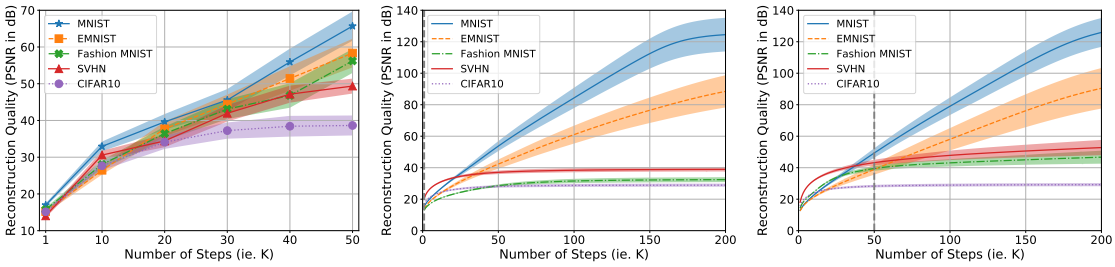
(a) MNIST **(b)** CIFAR10

Figure 7.4: Test results on shifted/flipped/rotated images using the reference learned on upright-centered (canonical) images. PSNR shown on top of images.



(a) Gaussian **(b)** Poisson

Figure 7.5: Reconstruction quality of the test images vs noise level of the measurements for different datasets. We learned the reference using noise-free measurements.



(a) Training K=Testing K **(b)** Training K=1 **(c)** Training K=50

Figure 7.6: Reconstruction PSNR vs the number of blocks (K) in the unrolled network at training and testing. (a) K is same for training and testing (shaded region shows ± 0.25 times `std` of PSNR). (b) $K = 1$ and (c) $K = 10$, but tested using different K.

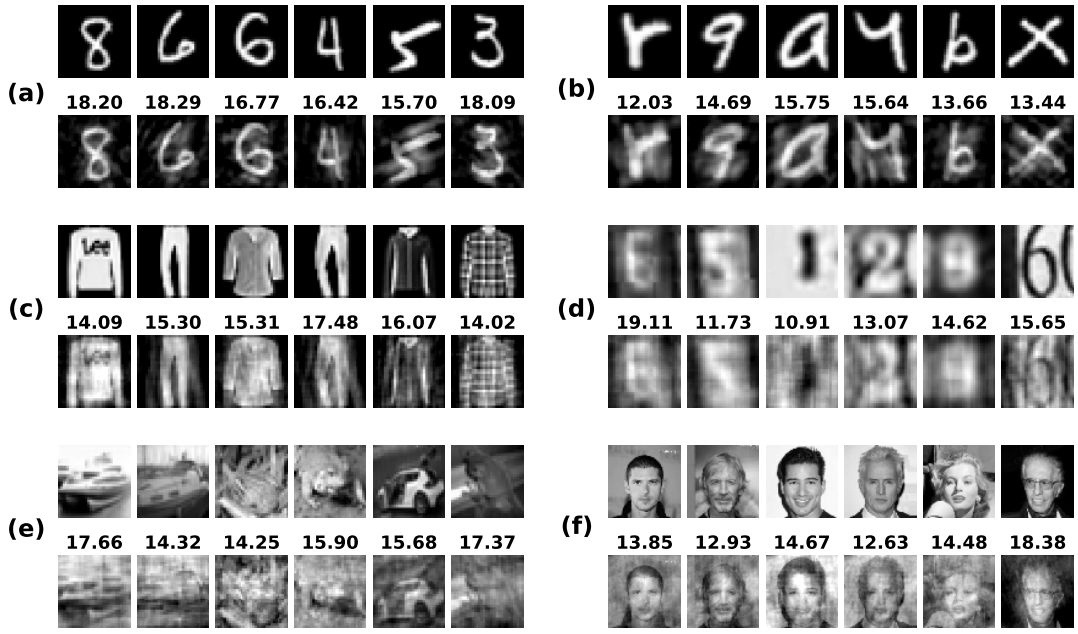
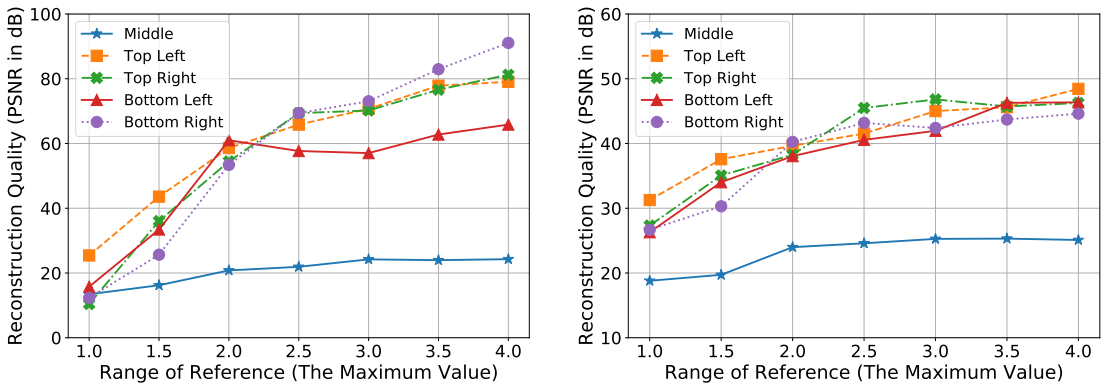


Figure 7.7: Single step reconstruction with reference in range $[0, 1]$. Each of the **6 sets (a)-(f)** has the the ground truth in the first row. Second row is the reconstruction (PSNR values on top).



(a) MNIST **(b)** CIFAR10

Figure 7.8: Performance of our method if the reference is an 8×8 block placed at different positions. Fixing the minimum value at 0, we increased the maximum value of the reference we learn. We observe that the small reference placed in the corners performs better than the ones placed in the center.

Chapter 8

Data-driven Illumination Patterns for Coded Diffraction Imaging

8.1 Introduction

The problem of signal recovery from nonlinear measurements arises in various imaging and signal processing tasks [424, 327, 67]. Conventional methods for solving such inverse problems use iterative methods to recover the signal from given measurements. In this paper, we present a framework to optimize the measurement parameters to improve the quality of signals recovered by the given iterative method. In particular, we learn illumination patterns to recover the signal from coded diffraction patterns (CDP) using a fixed-cost alternating minimization method.

We can model the sensor measurements for coded diffraction imaging as follows. Let us denote the signal of interest as $x \in \mathbb{R}^n$ or $\mathbb{C}^n$ that is modulated by T illumination patterns $D = \{d_1, \ldots, d_T\}$, where $d_t \in \mathbb{R}^n$ or $\mathbb{C}^n$. The amplitude of sensor measurements for t^{th} illumination

pattern can be written as

$$y_t = |\mathcal{F}(d_t \odot x)|, \tag{8.1}$$

where $\mathcal{F}$ denotes the Fourier transform operator and $\odot$ denotes an element-wise product. We note that real sensor measurements are proportional to the intensity of the incoming signal (i.e., square of the Fourier transform). In practice, however, solving the inverse problem with (non-square) amplitude measurements provides better results [523, 339]; therefore, we use the amplitude measurements throughout this paper.

To recover the signal x from the nonlinear measurements, we can solve the following optimization problem:

$$\min_x \sum_{t=1}^{T} \|y_t - |\mathcal{F}(d_t \odot x)|\|_2^2. \tag{8.2}$$

In recent years, a number of iterative algorithms have been proposed for solving the problem in (8.2), which includes lifting-based convex methods, alternating minimization-based nonconvex methods, and greedy methods [69, 167, 360, 215].

Our goal is to learn a set of illumination patterns to optimize the recovery of an alternating minimization (AltMin) algorithm for solving the problem in (8.2). The AltMin method can be viewed as an unrolled gradient descent network, where we fix the steps at every iteration and the total number of iterations for AltMin. One forward pass through the unrolled network is equivalent to K iterations of the AltMin algorithm. We can increase or decrease the number of iterations for better accuracy or faster run-time. To minimize the computational complexity of the recovery algorithm, we keep the total number of iterations small (e.g., $K = 50$). At the training stage, we optimize over the illumination patterns to minimize the error between the AltMin outputs after K iterations and the ground truth training images. At the test time, we solve the problem in (8.2) us-

ing K AltMin iteration with the learned illumination patterns (equivalent to one forward pass). We evaluated our method on different image datasets and compared against existing methods for coded diffraction imaging. We demonstrate that our proposed method of designing illumination patterns for a fixed-cost algorithm outperforms existing methods both in terms of accuracy and speed.

The key contributions of this paper are summarized as follows.

- We learned illumination patterns for a non-linear inverse problem (coded diffraction imaging) using unrolled network formulation of a classical AltMin method.
- We showed that with our designed patterns and unrolled AltMin method outperform computationally complex algorithms and provide superior image reconstruction.
- Our algorithm requires only a small number of training images to learn the illumination patterns. It is crucial in applications because finding training samples is difficult in practice.
- Our learned illumination patterns can also help other algorithms achieve higher performance even though they are not used for training.

8.2 Related Work

Phase Retrieval and Coded Diffraction Patterns: Coded diffraction imaging is a physically realistic setup of Fourier phase retrieval problem, where we can first modulate signal of interest and then collect the intensity measurements [67, 84]. The presence of modulation patterns makes this a more tractable problem compared to classical Fourier phase retrieval [67]. The algorithms for solving phase retrieval problem can be broadly divided into non-convex and convex methods. Amplitude flow [476], Wirtinger flow [68], alternating minimization (AltMin) [360] are recent methods that solve the non-convex problem. Convex methods usually lift the nonconvex problem of signal

recovery from quadratic measurements into a convex problem of low-rank matrix recovery from linear measurements [69, 429]. The PhaseLift algorithm [69] and its variations [167, 67] can be considered under this class. Other algorithms, such as PhaseMax [159] and PhaseLin [158], use convex relaxation to solve non-convex phase retrieval problem without lifting the problem to a higher dimension.

Data-Driven Approaches for Phase Retrieval: A number of papers have recently explored the idea of replacing the classical (hand-designed) signal priors with deep generative priors for solving inverse problems [52, 460, 464]. Another growing trend is to apply deep learning to solve inverse problems (including phase retrieval) in an end-to-end manner, where deep networks are trained to learn a mapping from sensor measurements to the signal of interest using a large number of measurement-signal pairs.

While our method is partially driven by data, our goal is not to learn a signal prior or a mapping from measurements to signal. We use data to learn the illumination patterns for a fixed recovery algorithm. The number of training images required by our method is extremely small (128 images only). Furthermore, the patterns we learn on one class of images provide good results on other types of images. Apart from the great flexibility, our method uses a well-defined AltMin routine, where we know exact steps for every iteration as opposed to the black-box deep models.

Unrolled Network for Inverse Problem: Iterative methods for solving the inverse problems, such as AltMin or other first-order methods, can be represented as unrolled networks. Every layer of such a network performs the same steps as a single iteration of the original method [53, 296]. Some parameters of the iterative steps can be learned from data (e.g., step size, denoiser, or threshold parameters) but the basic structure and physical forward model are kept intact. In our recent work

[213], we used the idea of unrolled network to solve phase retrieval problem from the holographic measurements.

Learn to Sense: Deep learning methods have also been recently used to design the sensing system; especially in the context of compressive sensing and computational imaging [498, 39]. The main objective, similar to ours, is to select sensor parameters to recover best possible signal/image from the sensor measurements. This may involve selection of samples/frames or illumination patters as we discuss in this paper. In contrast to most of the existing methods that learn a deep network to solve the inverse problem, our method uses a predefined iterative method as an unrolled network while learning the illumination patterns using a small number of training images. In principle, the sensor can be treated as the first layer of the network with some physical constraints on the parameters [247]. The method in [247] uses unrolled network to learn the sensing parameters for quantitative phase imaging problem under the "weak object approximation". This approximation turns the original nonlinear problem into a linear inverse problem. In our setup, we do not make any such assumptions on target object and solve the nonlinear inverse problem.

8.3 Proposed Method

We use N training images $(x_1, \ldots, x_N)$ to learn T illumination patterns that provide best reconstruction using a predefined (iterative) phase retrieval algorithm. Furthermore, to ensure that the illumination patterns are physically realizable, we constrain their values to be in the range $[0, 1]$. We use a sigmoid function over unconstrained parameters $\Theta = \{\theta_1, \ldots, \theta_T\}$ to define the illumination patterns; that is, $d_t = \text{sigmoid}(\theta_t)$ for all $t = 1, \ldots, T$.

Our proposed method for learning illumination patterns can be divided into two parts: The first (inner) part involves solving the phase retrieval problem with given coded diffraction patterns using AltMin as an unrolled network; Second part is updating the illumination patterns based on backpropagating the image reconstruction loss. These two parts provide optimized image reconstruction and illumination patterns. Pseudocodes for both parts are listed in Algorithms 7,8.

Phase retrieval as alternating minimization (AltMin): Given measurements $Y = \{y_1, \ldots, y_T\}$ and illumination patterns $D = \{d_1, \ldots, d_T\}$, we seek to solve the CDP phase retrieval problem by minimizing the loss function defined in (8.2) as

$$L_x = \frac{1}{T} \sum_{t=1}^{T} \| y_t - |\mathcal{F}(d_t \odot x)| \|_2^2. \tag{8.3}$$

Even though the loss function in (8.3) is nonconvex and nonsmooth with respect to x, we can minimize it using the well-known alternating minimization (AltMin) with gradient descent [360, 531]. We define a new variable for the estimated phase of linear measurements as $p_t = \text{phase}(\mathcal{F}(d_t \odot x))$ and reformulate the loss function in (8.3) into

$$L_{x,p} = \frac{1}{T} \sum_{t=1}^{T} \| p_t \odot y_t - \mathcal{F}(d_t \odot x) \|_2^2. \tag{8.4}$$

The gradient with respect to x can be computed as

$$\nabla_x L_{x,p} = \frac{2}{T} \sum_{t=1}^{T} |d_t|^2 \odot x - d_t^* \odot \mathcal{F}^*(p_t \odot y_t), \tag{8.5}$$

where $\mathcal{F}^*$ denotes the inverse Fourier transform and d_t^* is the conjugate of pattern d_t. We can update the estimate at every iteration as

$$x^k = x^{k-1} - \alpha_{k-1} \nabla_x L_{x,p}, \tag{8.6}$$

where α_{k-1} denotes the step size. Another way is to directly solve for x^k such that $\nabla_x L_{x,p} = 0$. The closed-form solution is

$$x^k = (\sum_{t=1}^{T} |d_t|^2)^{-1} \odot [\sum_{t=1}^{T} d_t^* \odot \mathcal{F}^*(p_t^{k-1} \odot y_t)]. \tag{8.7}$$

We compared these 2 strategies and found that single-step gradient descent tends to work well in practice and the closed-form solution does not show advantage over the single-step gradient descent. In our implementation, we used the former strategy (Algorithm 8) and fixed a step size α for all iterations. The unrolled network has K layers that implement K iterations of the gradient descent, and the final estimate is denoted as x^K.

Choice of initialization is important, and our method can handle different types of initialization. Zero initialization, where every pixel of the initial guess of x^0 is 0, is the simplest and cost-free method. Many recent phase retrieval algorithms [68, 98, 531, 24] use spectral initialization, which tries to find a good initial estimate. However, it requires computing the principal eigenvector of the following positive semidefinite matrix, $\sum_{t=1}^{T} \text{diag}(d_t^*)\mathcal{F}^*\text{diag}(|y_t|^2)\mathcal{F}\text{diag}(d_t)$. In our experiments, we observed that spectral initialization does not provide a significant improvement in terms of image reconstruction, and that our algorithm can perform very well using the overhead-free zero initialization.

Learning illumination patterns: To learn a set of illumination patterns that provide the best reconstruction with the predefined iterative method (or the unrolled network), we seek to minimize the difference between the original training images and their estimates. In this regard, we minimize the following quadratic loss function with respect to Θ:

$$L_\Theta = \sum_{n=1}^{N} \|x_n - x_n^K(\Theta)\|_2^2, \tag{8.8}$$

where $x_n^K(\Theta)$ denotes the `solveCDP` estimate of nth training image for the given values of Θ. Note that for given values of Θ, we can define illumination patterns as $d_t = \text{sigmoid}(\theta_t)$ and sensor measurements for x_n as $y_t^n = |\mathcal{F}(d_t \odot x_n)|$ for $t = 1, \ldots, T$ and $n = 1, \ldots, N$. We use Adam optimizer in PyTorch [253, 377] to minimize the loss function in (8.8). A summary of the algorithm for learning the illumination patterns is also listed in Algorithm 7.

8.4 Experiments

Datasets. We used MNIST digits and CelebA datasets for training and testing in our experiments. We used 128 images from each of the datasets for training and another 1000 images for testing. Images in CelebA dataset have 218×178 pixels, we first converted all the images to grayscale, cropped 178×178 region in the center, and resized to 200×200. Furthermore, we report the performance of our method on images used in [339] in Fig. 8.2.

Measurements. We used the amplitude of the 2D Fourier transform of the images modulated with T illumination patterns as the measurements. Unless otherwise mentioned, we used noiseless measurements.

Computing platform. We performed all the experiments using a computer equipped with Intel Core i7-8700 CPU and NVIDIA TITAN Xp GPU.

8.4.1 Setup and hyper-parameter search

The hyper-parameters include the number of iterations (K), step size α, and the number of training samples N. We set the default value of $K = 50$, but K can be adjusted as a trade-off between better reconstruction quality and shorter run time. We tested all methods for $T =$

Table 8.1: PSNR (`mean` $\pm$ `std`) for random and learned illumination patterns tested on different datasets.

Dataset	2 Illumination Patterns		3 Illumination Patterns		4 Illumination Patterns		8 Illumination Patterns	
	Random	Learned	Random	Learned	Random	Learned	Random	Learned
MNIST	14 ± 6	28 ± 9	20 ± 11	75 ± 19	32 ± 14	102 ± 10	61 ± 19	113 ± 11
CelebA	13 ± 2	19 ± 3	14 ± 4	28 ± 2	23 ± 5	81 ± 4	43 ± 8	98 ± 15

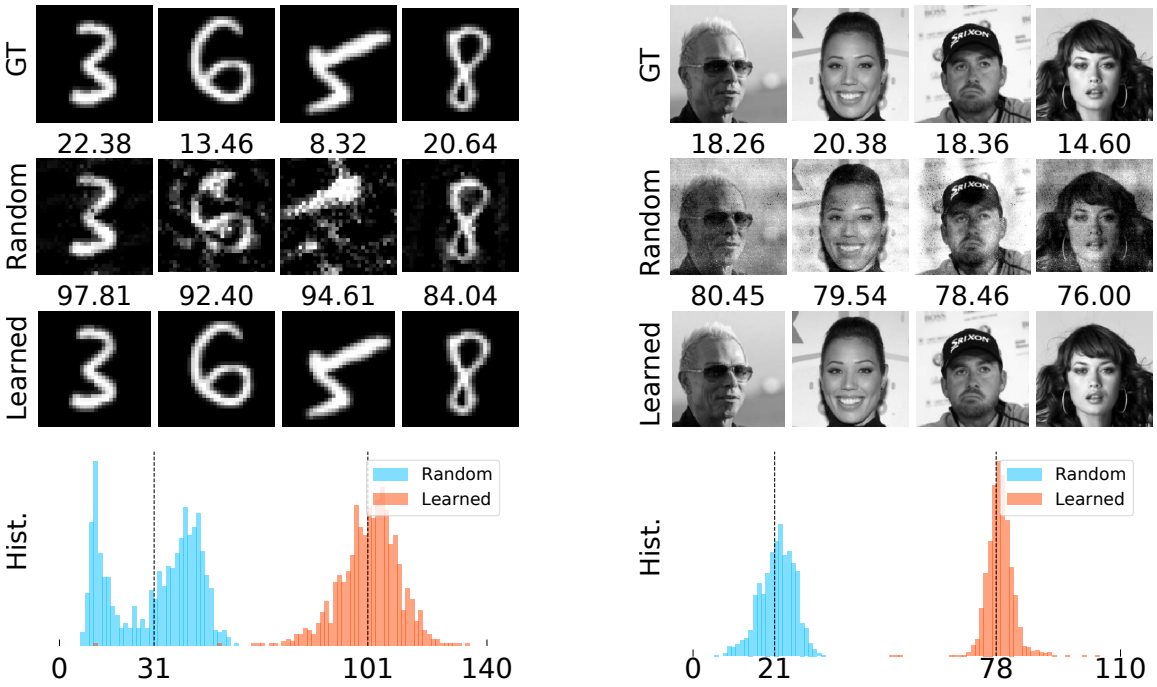
$\{2, 3, 4, 8\}$ to evaluate cases where signal recovery is hard, moderate, and easy. Through grid search, we found that it provides the best results over all datasets when $\alpha = 4/T$.

8.4.2 Comparison between random and learned patterns

To demonstrate the advantages of our learned illumination patterns, we compare the performance of learned and random illumination patterns on five different datasets. We learn a set of $T = \{2, 3, 4, 8\}$ illumination patterns on 128 training images from a dataset and test them on 1000 test images from the same dataset. For random patterns, we draw T independent patterns from Uniform(0,1) distribution and test their performance on the same 1000 samples that we used for the learned case. We repeat this process 30 times and choose the best result to compare with the results for the learned illumination patterns. The average PSNR over all 1000 test image reconstructions is presented in Table 8.1, which shows that the learned illumination patterns perform significantly better than the random patterns for all values of T. In addition to that, we can observe a transition in the performance for $T = 3$, where random patterns provide poor quality reconstructions and learned

patterns provide reasonably high quality reconstructions. Furthermore, the learned patterns provide very high quality reconstructions for $T \geq 4$.

To highlight this effect, we show a small set of reconstructed images and histograms of PSNRs of all the reconstructed images from learned and random illumination patterns in Fig. 8.1 for $T = 4$ patterns. The result suggests that the learned illumination patterns demonstrate consistently better performance compared to random illumination patterns.



(a) MNIST **(b)** CelebA

Figure 8.1: Selected ground truth (GT) images, corresponding reconstructed images using random and learned illumination patterns. PSNR is shown on top of every reconstruction. Below each dataset, we show the histograms of the PSNRs of all images with random patterns (shown in blue) and learned patterns (shown in orange). The dashed vertical line indicates the mean of all PSNRs.

Table 8.2: Reconstruction PSNR (`mean` $\pm$ `std`) of different algorithms using random patterns and our learned patterns. The number of patterns is 4 in each case. Here we round the PSNR values to integers to fit the width of the page.

*For Deep Model [337] experiments, patterns are normalized to $[-1, 1]$ range. **For Deep Model, the image size for CelebA generator is 64×64.

	MNIST		CelebA	
	Random	Learned	Random	Learned
HIO [149]	16 ± 9	37 ± 19	38 ± 5	102 ± 5
GS [157]	16 ± 9	37 ± 19	38 ± 4	102 ± 5
WirtFlow [68]	22 ± 16	48 ± 25	20 ± 2	39 ± 3
AmpFlow [98]	42 ± 32	74 ± 48	42 ± 8	138 ± 11
PhaseMax [24]	14 ± 4	24 ± 8	32 ± 2	148 ± 2
Deep Model [337]*	31 ± 2	32 ± 3	22 ± 3**	23 ± 2**
Ours - K=50	32 ± 14	102 ± 10	23 ± 5	81 ± 4
Ours - K=100	51 ± 19	186 ± 15	33 ± 4	132 ± 7

8.4.3 Comparision with existing methods

We compare our method with various existing methods using different datasets. These existing methods fall into 4 categories:

- Hybrid input output (HIO) [149] and Gerchberg-Saxton (GS) [157] (alternating minimization methods)
- Wirtinger Flow [68] and Amplitude Flow [98] (non-convex, gradient descent-based methods)

- PhaseMax [24] (a convex method)
- Deep S^3PR [337] (deep model-based method).

We compare the performance of our method with these methods in terms of reconstruction quality.

For algorithms in [149, 157, 68, 98, 24], we used PhasePack [84] package with default spectral initialization. In our comparison, we used $T = 4$ illumination patterns and in value range of $[0, 1]$. To make a fair comparison between our models in reconstruction quality, we set the error tolerance ($\texttt{tol} = 10^{-6}$) and run each algorithm till convergence.

For deep generative models, we used a modified version of the publicly available code and DCGAN model for MNIST from [337] and trained our DCGAN model for CelebA. This method is noticeably time-consuming because it optimizes over the latent vector for the deep model and uses 2000 iterations for each image where each iteration requires a forward and backward pass through the deep model. The reconstruction results for the Deep Model also directly depend on the quality of the trained generative models. In our experiments, we were not able to generate images with PSNR higher than 30dB using the generative models.

For the case of Random illumination, we selected the best PSNR from 5 independent trials and report the average computation time for each experiment. In all the cases, we tuned the parameters that provide best results.

The reconstruction PSNR (in dB) is reported in Table 8.2. We observe that our proposed method with learned patterns performs significantly better than all other algorithms in reconstruction quality.

An interesting attribute of our learned patterns is that they can be used with different algorithms. We observe in Table 8.2 that our learned patterns provide better results compared to

Random patterns with all the phase retrieval algorithms, even though the patterns were not optimized for those algorithms.

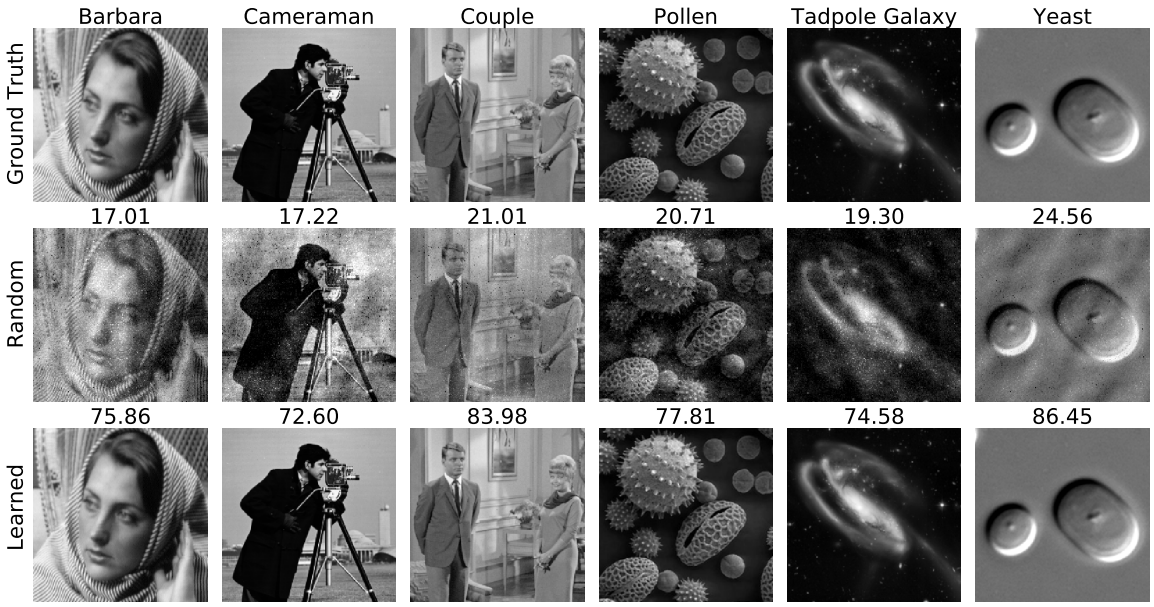


Figure 8.2: First Row: Ground truth images from image processing standard test datasets. **Second Row:** Reconstruction using random illumination patterns with uniform random distribution [0, 1] (we selected $T = 4$ patterns that provided best results on celebA test images in **30 trials**). PSNR numbers are shown on the top of reconstructed images. **Third Row:** Reconstruction using the patterns trained on celebA dataset. Each image has 200×200 pixels and the number of illumination patterns is $T = 4$.

8.4.4 Generalization of learned patterns on different datasets

To explore the generalizability of our learned illumination patterns, we use patterns learned on one dataset to recover several classical images. Some results are shown in Fig. 8.2. We used illumination patterns learned on 128 celebA images, but we can see that the learned illumination

patterns perform better than the randomly chosen illumination patterns for classical images which supports the generalizability of our learned illumination patterns.

8.5 Conclusion

We presented a framework to learn the illumination patterns for coded diffraction imaging by formulating an iterative phase retrieval algorithm as a fixed unrolled network. It only takes a small number of training images to achieve near-perfect reconstruction whereas random patterns fail. In addition, the learning process of our illumination patterns is highly data efficient and requires a small number of training samples. The learned patterns generalize to different datasets and algorithms that were not used during training.

Algorithm 7 Learning Illumination Patterns

Input:

Training set X with N images $X = \{x_1, \ldots, x_N\}$

Output:

Learned illumination patterns $D = \{d_1, \ldots, d_T \mid d_t = \text{sigmoid}(\theta_t)\}$

Initialize optimization variables for T patterns as $\Theta = \{\theta_1, \ldots, \theta_T\}$ from a uniform distribution ▷ *uniform initialization*

for epoch $= 1, 2, ..., M$ **do**

  Generate illumination patterns $d_t = \text{sigmoid}(\theta_t)$ for all t ▷ *generate patterns*

  for $n = 1, 2, ..., N$ **do**

    Compute measurements $Y^n = \{y_1^n, \ldots, y_T^n \mid y_t^n = |\mathcal{F}(d_t \odot x_n)|\}$ ▷ *compute measurements*

    Reconstruct image $x_n^K(\Theta) \leftarrow \texttt{solveCDP}(Y^n, D)$ ▷ *reconstruct image using CDP*

  end for

  Compute loss $L_\Theta = \sum_{n=1}^{N} \|x_n - x_n^K(\Theta)\|_2^2$ ▷ *compute loss*

  Update $\Theta \leftarrow \Theta - \beta \nabla_\Theta L_\Theta$ with stepsize β ▷ *gradient descent update*

end for

return Learned illumination patterns $D = \{d_1, \ldots, d_T \mid d_t = \text{sigmoid}(\theta_t)\}$

Algorithm 8 `solveCDP` (Y, D) via Alternating Minimization

Input: Measurements $Y = \{y_1, \ldots, y_t\}$ and illumination patterns $D = \{d_1, \ldots, d_T\}$.

Output: Estimated signal x^K.

1: **Initialization:** Zero initialization of estimate x^0.

2: **for** $k = 1, 2, ..., K$ **do**

3: $p_t^{k-1} \leftarrow \text{phase}(\mathcal{F}(d_t \odot x^{k-1}))$ for all t.

4: $\nabla_x L_{x,p} = \frac{2}{T} \sum_{t=1}^{T} [|d_t|^2 \odot x^{k-1} - d_t^* \odot \mathcal{F}^*(p_t^{k-1} \odot y_t)]$

5: $x^k \leftarrow x^{k-1} - \alpha \nabla_x L_{x,p}$

6: Project x^k onto feasible range.

7: **end for**

8: **return** Estimated signal x^K.

Chapter 9

Conclusions

To summarize, my research has focused on identifying and mitigating vulnerabilities in neural networks, with a particular emphasis on enhancing AI trustworthiness and preserving user privacy. We have developed efficient black-box attacks, explored context-aware adversarial methods, proposed innovative techniques for facial de-identification, and developed methods to enhance computational imaging frameworks.

The contributions of this work are significant in ensuring that AI systems are robust and reliable, especially as they become more prevalent in critical domains. The implications of these findings extend beyond academic research, offering practical solutions for deploying AI technologies safely and responsibly.

Looking ahead, there are numerous opportunities for future research, including further refinement of attack and defense techniques and exploring their applications in emerging AI domains.

Large-Scale Real-World Impacts of AI System Vulnerabilities

Systematically assessing the potential of AI systems to cause large-scale real-world disasters is essential as models grow more capable and sophisticated. I plan to generalize my ensemble attacks across textual, audio, visual, and multimodal systems, particularly concentrating on large language models (LLMs), generative networks, and models that leverage contextual reasoning. Collaborating actively with industry leaders will be crucial to translating these academic advancements into real-world impacts and ensuring the practical deployment of trustworthy AI.

Development of Reliable and Responsible AI Systems

The construction of robust AI systems involves not just addressing current vulnerabilities but also anticipating future challenges by developing intelligent context-aware systems and integrating physics-based inductive biases:

- I will develop defenses leveraging language models for contextual input vetting. By specializing models on conversational data and adversaries, fine-tuning can strengthen anomaly detection. Inputs undergo consistency checks using conditional likelihoods to discern idiosyncrasies from malicious attempts. Quantifying reliability based on detection rates and accuracy enables standardized vetting to fortify deployed systems.

- Additionally, I will quantify the adversarial resilience physics-based priors provide in computational imaging models [62, 213, 217]. By benchmarking attack success rates across modalities and data-driven models, I can demonstrate cases where physical constraints confer reliability advantages. Validating how factors like noise and occlusion impact difficulty highlights advantages of hybrid physics-architecture integration. Incorporating differential privacy can

further ensure confidentiality. My vision centers on multifaceted research enabling robust and responsible intelligent systems.

Building and Securing AI Agents in Reinforcement Learning Environments

A compelling new avenue I am excited to investigate is the creation of resilient AI agents operating within reinforcement learning environments, particularly under conditions marked by significant uncertainty and variability. This research will explore the hypothesis that diverse, low-fidelity simulations can facilitate a more effective and rapid transfer of learned behaviors to physical systems, thereby dramatically narrowing the simulation-to-reality gap. By employing this strategy, we aim to rigorously test the theory that simplified models can expedite the adaptation of autonomous systems in dynamic, real-world settings, potentially revolutionizing our approach to AI training and deployment. This could pave the way for AI systems that are not only smarter and more efficient but are also better equipped to handle the unpredictable nuances of the real world.

Bibliography

[1] Health insurance portability and accountability act. U.S. Department of Health and Human Services, 2003. 45 CFR Parts 160, 162, and 164.

[2] California consumer privacy act. California Legislative Information, 2018. Cal. Civ. Code §1798.100 et seq.

[3] Personal information protection law. National People's Congress of the People's Republic of China, 2021.

[4] InsightFace: 2D and 3D Face Analysis Project, 2022. https://github.com/deepinsight/insightface.

[5] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[6] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM CCS*, 2016.

[7] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments. *arXiv:2203.03339*, 2022.

[8] Ayush Agarwal, Pratik Chattopadhyay, and Lipo Wang. Privacy preservation through facial de-identification with simultaneous emotion preservation. *SIVP*, 15(5), 2021.

[9] Michal Aharon, Michael Elad, and Alfred Bruckstein. rmk-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[10] Abhishek Aich, Akash Gupta, Rameswar Panda, Rakib Hyder, M. Salman Asif, and Amit K. Roy-Chowdhury. Non-adversarial video synthesis with learned priors, 2020.

[11] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[12] Abdullah Al-Dujaili and Una-May O'Reilly. There are no bit parts for sign bits in black-box attacks. *arXiv preprint arXiv:1902.06894*, 2019.

[13] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020.

[14] Fahimeh Arab and M Salman Asif. Fourier phase retrieval with arbitrary reference signal. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1479–1483. IEEE, 2020.

[15] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[16] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018.

[17] M.S. Asif and C. Hegde. Phase retrieval for signals in union of subspaces. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 356–359. IEEE, 2018.

[18] Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Trans. on Comp. Imaging*, 2017.

[19] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, pages 274–283, 2018.

[20] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[21] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[23] S. Bahmani and J. Romberg. Efficient compressive phase retrieval with constrained sensing vectors. pages 523–531, 2015.

[24] S. Bahmani and J. Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. In *Artificial Intelligence and Statistics*, pages 252–260, 2017.

[25] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *International Joint Conference on Artificial Intelligence*, 2021.

[26] Radu Balan. On signal reconstruction from its spectrogram. In *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–4. IEEE, 2010.

[27] Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

[28] Jawadul H Bappy, Sujoy Paul, and Amit K Roy-Chowdhury. Online adaptation for joint scene and object classification. In *European Conference on Computer Vision*, pages 227–243. Springer, 2016.

[29] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, 2019.

[30] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, 2010.

[31] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *CVPR*, 2023.

[32] D.A. Barmherzig, J. Sun, P. Li, T.J. Lane, and E. Candès. Holographic phase retrieval and reference design. *Inverse Problems*, 2019.

[33] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7412–7420, 2019.

[34] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[35] Heinz H Bauschke, Patrick L Combettes, and D Russell Luke. Phase retrieval, error reduction algorithm, and fienup variants: a view from convex optimization. *JOSA A*, 19(7):1334–1345, 2002.

[36] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[37] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2874–2883, 2016.

[38] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[39] A. W. Bergman, D. B. Lindell, and G. Wetzstein. Deep Adaptive LiDAR: End-to-end Optimization of Sampling and Depth Completion at Low Sampling Rates. *Proc. IEEE ICCP*, 2020.

[40] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[41] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. *arXiv preprint arXiv:1704.02654*, 2, 2017.

[42] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *European Conference on Computer Vision*, pages 158–174. Springer, 2018.

[43] A. Bhandari, F. Krahmer, and R. Raskar. On unlimited sampling. pages 31–35, 2017.

[44] Bir Bhanu, Ajay Kumar, et al. *Deep learning for biometrics*, volume 7. Springer, 2017.

[45] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[46] Battista Biggio, Giorgio Fumera, and Fabio Roli. Pattern recognition systems under attack: Design issues and research challenges. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(07):1460002, 2014.

[47] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

[48] David SC Biggs. 3d deconvolution microscopy. *Current Protocols in Cytometry*, pages 12–19, 2010.

[49] J. Bioucas-Dias and G. Valadão. Phase unwrapping via graph cuts. 16(3), 2007.

[50] J. Bioucas-Dias and G. Valadao. Phase unwrapping via graph cuts. *IEEE Trans. Image Proc.*, 16(3):698–709, 2007.

[51] P. Bojanowski, A. Joulin, D. Lopez-Paz, and A. Szlam. Optimizing the latent space of generative networks. 2018.

[52] A. Bora, A. Jalal, E. Price, and A. Dimakis. Compressed sensing using generative models. 2017.

[53] Emrah Bostan, Ulugbek S Kamilov, and Laura Waller. Learning-based image reconstruction via parallel proximal algorithm. *IEEE Signal Processing Letters*, 25(7):989–993, 2018.

[54] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *CSCW*, 2000.

[55] Google Brain. Neurips 2017: Targeted adversarial attack. https://www.kaggle.com/competitions/nips-2017-targeted-adversarial-attack/data, 2017. [On Kaggle].

[56] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *International Conference on Learning Representations*, 2018.

[57] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.

[58] Darren Burke and Danielle Sulikowski. The evolution of holistic processing of faces. *Frontiers in psychology*, 4:11, 2013.

[59] T. Cai, X. Li, Z. Ma, et al. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. 44(5):2221–2251, 2016.

[60] Zikui Cai, Zhongpai Gao, Benjamin Planche, Meng Zheng, Terrence Chen, M Salman Asif, and Ziyan Wu. Disguise without disruption: Utility-preserving face de-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 918–926, 2024. https://arxiv.org/abs/2303.13269.

[61] Zikui Cai, Rakib Hyder, and M Salman Asif. Learning illumination patterns for coded diffraction phase retrieval. *arXiv preprint arXiv:2006.04199*, 2020.

[62] Zikui Cai, Rakib Hyder, and M Salman Asif. Data-driven illumination patterns for coded diffraction imaging. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2818–2822. IEEE, 2021.

[63] Zikui Cai, Shantanu Rane, Alejandro E Brito, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and M Salman Asif. Zero-query transfer attacks on context-aware object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15024–15034, 2022.

[64] Zikui Cai, Chengyu Song, Srikanth Krishnamurthy, Amit Roy-Chowdhury, and M. Salman Asif. Blackbox attacks via surrogate ensemble search. In *Advances in Neural Information Processing Systems*, 2022.

[65] Zikui Cai, Yaoteng Tan, and M Salman Asif. Ensemble-based blackbox attacks on dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4045–4055, 2023. https://arxiv.org/abs/2303.14304.

[66] Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and M. Salman Asif. Context-aware transfer attacks for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 149–157, 2022.

[67] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. 39(2):277–299, 2015.

[68] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: theory and algorithms. 61(4):1985–2007, 2015.

[69] E. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. 66(8):1241–1274, 2013.

[70] Emmanuel J Candès et al. Compressive sampling. In *Proceedings of the international congress of mathematicians*, volume 3, pages 1433–1452. Madrid, Spain, 2006.

[71] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[72] Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, and Li Song. Personalized and invertible face de-identification by disentangled identity information manipulation. In *ICCV*, 2021.

[73] Nan Cao, Chaoguang Lin, Qiuhan Zhu, Yu-Ru Lin, Xian Teng, and Xidao Wen. Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data. *IEEE transactions on visualization and computer graphics*, 24(1):23–33, 2017.

[74] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[75] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[76] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[77] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *USENIX Security Symposium*, pages 513–530, 2016.

[78] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[79] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[80] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[81] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[82] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1):89–97, 2004.

[83] Tony F Chan, Jianhong Shen, and Hao-Min Zhou. Total variation wavelet inpainting. *Journal of Mathematical imaging and Vision*, 25(1):107–125, 2006.

[84] Rohan Chandra, Ziyuan Zhong, Justin Hontz, Val McCulloch, Christoph Studer, and Tom Goldstein. Phasepack: A phase retrieval library. *Asilomar Conference on Signals, Systems, and Computers*, 2017.

[85] H. Chang, Y. Lou, M.K. Ng, and T. Zeng. Phase retrieval from incomplete magnitude information via total variation regularization. *SIAM Journal on Scientific Computing*, 38(6):A3672–A3695, 2016.

[86] Zhaohui Che, Ali Borji, Guangtao Zhai, Suiyi Ling, Jing Li, and Patrick Le Callet. A new ensemble adversarial attack powered by long-term gradient memories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3405–3413, 2020.

[87] Jia-Wei Chen, Li-Ju Chen, Chia-Mu Yu, and Chun-Shien Lu. Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics. In *CVPR*, 2021.

[88] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

[89] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection. https://github.com/open-mmlab/mmdetection, 2019. [Apache License 2.0].

[90] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[91] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[92] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[93] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM International Conference on Multimedia*, 2020.

[94] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[95] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018.

[96] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. pages 2172–2180, 2016.

[97] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4086–4096, 2017.

[98] Y. Chen and E. Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. pages 739–747, 2015.

[99] Z. Chen, G. Jagatap, S. Nayer, C. Hegde, and N. Vaswani. Low rank fourier ptychography. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6538–6542, April 2018.

[100] Zhe Chen, Shaoli Huang, and Dacheng Tao. Context refinement for object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 71–86, 2018.

[101] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *Advances in Neural Information Processing Systems*, volume 32, pages 10934–10944, 2019.

[102] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 514–532. Springer, 2022.

[103] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[104] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. A tree-based context model for object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 34(2):240–252, 2011.

[105] Ka-Ho Chow, Ling Liu, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu. Tog: Targeted adversarial objectness gradient attacks on real-time object detection systems. *arXiv preprint arXiv:2004.04320*, 2020.

[106] Ka-Ho Chow, Ling Liu, Margaret Loper, Juhyun Bae, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu. Adversarial objectness gradient attacks in real-time object detection systems. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 263–272. IEEE, 2020.

[107] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

[108] MMCV Contributors. MMCV: OpenMMLab computer vision foundation. https://github.com/open-mmlab/mmcv, 2018.

[109] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[110] John V Corbett. The pauli problem, state reconstruction and quantum-real numbers. *Reports on Mathematical Physics*, 1(57):53–68, 2006.

[111] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[112] Andrei Costin. Security of cctv and video surveillance systems: threats, vulnerabilities, attacks, and mitigations. In *Proceedings of the 6th International Workshop on Trustworthy Embedded Devices*, pages 45–54. ACM, 2016.

[113] William L Croft, Jörg-Rüdiger Sack, and Wei Shi. Obfuscation of images via differential privacy: From facial images to general images. *P2PNA*, 14(3), 2021.

[114] M. Cucuringu and H. Tyagi. On denoising modulo 1 samples of a function. 2018.

[115] M. Cucuringu and H. Tyagi. Provably robust estimation of modulo 1 samples of a smooth function with applications to phase unwrapping. 2018.

[116] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, 2021.

[117] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–204, 2018.

[118] Jester Dataset. Humans performing pre-defined hand actions. https://20bn.com/datasets/jester, 2016. [Online; accessed 30-April-2018].

[119] Jean-Marc Deltorn. Deep creations: Intellectual property and the automata. *Frontiers in Digital Humanities*, 4:3, 2017.

[120] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[121] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

[122] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[123] Chaitanya Desai, Deva Ramanan, and Charless C Fowlkes. Discriminative models for multi-class object layout. *International Journal of Computer Vision*, 95, 2011.

[124] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6), 2012.

[125] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[126] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

[127] Steven Diamond, Vincent Sitzmann, Felix Heide, and Gordon Wetzstein. Unrolled optimization with deep priors. *arXiv preprint arXiv:1705.08041*, 2017.

[128] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1271–1278, 2009.

[129] C. Dong, C.C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[130] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

[131] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[132] David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995.

[133] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[134] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[135] Jiawei Du, Hu Zhang, Joey Tianyi Zhou, Yi Yang, and Jiashi Feng. Query-efficient meta attack to deep neural networks. In *International Conference on Learning Representations*, 2020.

[136] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *FOCS*. IEEE, 2013.

[137] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018.

[138] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 2014.

[139] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

[140] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[141] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. 2018.

[142] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[143] Sean Ryan Fanello, Ilaria Gori, Giorgio Metta, and Francesca Odone. One-shot learning for real-time action recognition. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 31–40. Springer, 2013.

[144] Maryam Fazel, E Candes, Benjamin Recht, and P Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2008.

[145] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[146] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[147] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 2010.

[148] C Fienup and J Dainty. Phase retrieval and image reconstruction for astronomy. *Image recovery: theory and application*, 231:275, 1987.

[149] J. R. Fienup. Reconstruction of an object from the modulus of its fourier transform. *Optics letters*, 3(1):27–29, 1978.

[150] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[151] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008.

[152] Homa Foroughi, Baharak Shakeri Aski, and Hamidreza Pourreza. Intelligent video surveillance for monitoring fall detection of elderly in home environments. In *Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on*, pages 219–224. IEEE, 2008.

[153] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM CCS*, 2015.

[154] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in google street view. In *ICCV*, 2009.

[155] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010.

[156] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[157] R. W. Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[158] A. S Ghods, Ra.and Lan, T. Goldstein, and C. Studer. Phaselin: Linear phase retrieval. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2018.

[159] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, 2018.

[160] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017.

[161] Robert A Gonsalves. Perspectives on phase retrieval and phase diversity in astronomy. In *Adaptive Optics Systems IV*, volume 9148, page 91482P. International Society for Optics and Photonics, 2014.

[162] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. pages 2672–2680, 2014.

[163] Ian Goodfellow. A research agenda: Dynamic models to defend against correlated attacks. *arXiv preprint arXiv:1903.06293*, 2019.

[164] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11), 2020.

[165] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[166] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 399–406, 2010.

[167] D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. 42(1):37–64, 2017.

[168] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. Integrating utility into face de-identification. In *International Workshop on Privacy Enhancing Technologies*. Springer, 2005.

[169] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. Face de-identification. In *Protecting privacy in video surveillance*. Springer, 2009.

[170] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Model-based face de-identification. In *CVPR workshop*, 2006.

[171] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial perturbations against deep neural networks for malware classification. *arXiv preprint arXiv:1606.04435*, 2016.

[172] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip Torr. Adversarial examples on segmentation models can be easy to transfer. *arXiv preprint arXiv:2111.11368*, 2021.

[173] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer, 2022.

[174] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 2017.

[175] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *ECCV*. Springer, 2020.

[176] M. Guizar-Sicairos and J.R. Fienup. Holography with extended reference by autocorrelation linear differential operation. *Optics express*, 15(26):17592–17612, 2007.

[177] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019.

[178] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2017.

[179] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

[180] P. Hand, O. Leong, and V. Voroninski. Phase retrieval under a generative prior. pages 9154–9164, 2018.

[181] Paul Hand and Vladislav Voroninski. Compressed sensing from phaseless gaussian measurements via linear programming in the natural parameter space. *arXiv preprint arXiv:1611.05985*, 2016.

[182] R. Harrison. Phase problem in crystallography. *JOSA a*, 10(5):1046–1055, 1993.

[183] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742. IEEE, 2016.

[184] Mahmudul Hasan, Sujoy Paul, Anastasios I Mourikis, and Amit K Roy-Chowdhury. Context-aware query selection for active learning in event recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[185] Mahmudul Hasan and Amit K Roy-Chowdhury. Context aware active learning of activity recognition models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4543–4551, 2015.

[186] K He, X Zhang, S Ren, et al. Spatial pyramid pooling in deep convolution networks or visual classification. In *ECCV*, 2014.

[187] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[188] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–597, 2019.

[189] R. Heckel and P. Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. 2018.

[190] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. *arXiv:2202.12555*, 2022.

[191] Dan Hendrycks and Kevin Gimpel. Early methods for detecting adversarial images. *arXiv preprint arXiv:1608.00530*, 2016.

[192] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

[193] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[194] Andrew Hollingworth. Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127(4):398, 1998.

[195] A. Hooper and H. Zebker. Phase unwrapping in three dimensions with application to InSAR time series. *J. of the Optical Soc. of America A*, 24(9):2737, 2007.

[196] Hossein Hosseini, Baicen Xiao, and Radha Poovendran. Google's cloud vision api is not robust to noise. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 101–105. IEEE, 2017.

[197] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[198] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018.

[199] Shengnan Hu, Yang Zhang, Sumit Laha, Ankit Sharma, and Hassan Foroosh. Cca: Exploring the possibility of contextual camouflage attack on object detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7647–7654. IEEE, 2021.

[200] Weiwei Hu and Ying Tan. Generating adversarial malware examples for black-box attacks based on gan. *arXiv preprint arXiv:1702.05983*, 2017.

[201] Chi-Hsuan Huang, Tsung-Han Lee, Lin-huang Chang, Jhih-Ren Lin, and Gwoboa Horng. Adversarial attacks on sdn-based deep learning ids system. In *International Conference on Mobile and Wireless Technology*, pages 181–191. Springer, 2018.

[202] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[203] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images*, 2008.

[204] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58. ACM, 2011.

[205] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4733–4742, 2019.

[206] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 4, 2017.

[207] Zhanchao Huang, Jianlin Wang, Xuesong Fu, Tao Yu, Yongqi Guo, and Rutong Wang. Dc-spp-yolo: Dense connection and spatial pyramid pooling based yolo for object detection. *Information Sciences*, 2020.

[208] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2019.

[209] Zhichao Huang and Tong Zhang. Tremba. https://github.com/TransEmbedBA/TREMBA, 2019. [No license provided].

[210] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[211] Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *WACV*, 2023.

[212] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *ISVC*. Springer, 2019.

[213] R. Hyder, Z. Cai, and M. S. Asif. Solving phase retrieval with a learned reference. 2020.

[214] R. Hyder, C. Hegde, and M.S. Asif. Fourier phase retrieval with side information using generative prior. IEEE, 2019.

[215] R. Hyder, Viraj S., C. Hegde, and M.S. Asif. Alternating phase projected gradient descent with generative priors for solving compressive phase retrieval. pages 7705–7709. IEEE, 2019.

[216] Rakib Hyder and M. Salman Asif. Generative models for low-dimensional video representation and reconstruction. *IEEE Transactions on Signal Processing*, 68:1688–1701, 2020.

[217] Rakib Hyder, Zikui Cai, and M Salman Asif. Learning to sense for coded diffraction imaging. *Sensors*, 22(24):9964, 2022.

[218] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[219] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *International Conference on Machine Learning*, pages 2137–2146, 2018.

[220] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018.

[221] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

[222] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7066–7074, 2019.

[223] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[224] L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk. Robust 1-Bit compressive sensing via binary stable embeddings of sparse vectors. 59(4):2082–2102, 2013.

[225] K. Jaganathan, S. Oymak, and B. Hassibi. Recovery of sparse 1-d signals from the magnitudes of their fourier transform. pages 1473–1477. IEEE, 2012.

[226] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. Phase retrieval: An overview of recent developments. *arXiv preprint arXiv:1510.07713*, 2015.

[227] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. Stft phase retrieval: Uniqueness guarantees and recovery algorithms. *IEEE Journal of selected topics in signal processing*, 10(4):770–781, 2016.

[228] G. Jagatap, Z. Chen, C. Hegde, and N. Vaswani. Sub-diffraction imaging using fourier ptychography and structured sparsity. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497, April 2018.

[229] G. Jagatap, Z. Chen, S. Nayer, C. Hegde, and N. Vaswani. Sample efficient fourier ptychography for structured data. *IEEE Transactions on Computational Imaging*, 6:344–357, 2020.

[230] G. Jagatap and C. Hegde. Fast, sample-efficient algorithms for structured phase retrieval. In *Advances in Neural Information Processing Systems*, pages 4917–4927, 2017.

[231] Gauri Jagatap and Chinmay Hegde. Algorithmic guarantees for inverse imaging with untrained network priors. In *Advances in Neural Information Processing Systems*, pages 14832–14842, 2019.

[232] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[233] Xiaojun Jia, Xingxing Wei, and Xiaochun Cao. Identifying and resisting adversarial videos using temporal consistency. *arXiv preprint arXiv:1909.04837*, 2019.

[234] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6084–6092, 2019.

[235] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 864–872, 2019.

[236] Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur. Mmm: Multi-stage multi-task learning for multi-choice reading comprehension. *arXiv preprint arXiv:1910.00458*, 2019.

[237] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *ICB*. IEEE, 2015.

[238] Barry L Kalman and Stan C Kwasny. Why tanh: choosing a sigmoidal function. In *Neural Networks, 1992. IJCNN., International Joint Conference on*, volume 4, pages 578–581. IEEE, 1992.

[239] U. Kamilov, V. Goyal, and S. Rangan. Message-passing de-quantization with applications to compressed sensing. *IEEE Trans. Sig. Proc.*, 60(12):6270–6281, 2012.

[240] Ulugbek S Kamilov and Hassan Mansour. Learning optimal nonlinearities for iterative thresholding algorithms. *IEEE Signal Processing Letters*, 23(5):747–751, 2016.

[241] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[242] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv:1710.10196*, 2017.

[243] Hirokatsu Kataoka, Yutaka Satoh, Yoshimitsu Aoki, Shoko Oikawa, and Yasuhiro Matsui. Temporal and fine-grained pedestrian action recognition on driving recorder database. *Sensors*, 18(2):627, 2018.

[244] Hirokatsu Kataoka, Teppei Suzuki, Shoko Oikawa, Yasuhiro Matsui, and Yutaka Satoh. Drive video analysis for the detection of traffic near-miss incidents. *arXiv preprint arXiv:1804.02555*, 2018.

[245] S. Kavusi and A. El Gamal. Quantitative study of high-dynamic-range image sensor architectures. In *Sensors and Camera Systems for Sci., Indust., and Digi. Photography Applications V*, volume 5301, pages 264–276. Intl. Soc. for Optics and Photonics, 2004.

[246] Michael Kellman, Emrah Bostan, Michael Chen, and Laura Waller. Data-driven design for fourier ptychographic microscopy. *International Conference for Computational Photography*, pages 1–8, 2019.

[247] Michael R Kellman, Emrah Bostan, Nicole A Repina, and Laura Waller. Physics-based learned design: optimized coded-illumination for quantitative phase imaging. *IEEE Transactions on Computational Imaging*, 5(3):344–353, 2019.

[248] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019.

[249] Brendan Kelly, Thomas P Matthews, and Mark A Anastasio. Deep learning-guided image reconstruction from incomplete data. *arXiv preprint arXiv:1709.00584*, 2017.

[250] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.

[251] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *CVPR*, 2022.

[252] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10, 2009.

[253] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[254] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

[255] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *NeurIPS*, 28, 2015.

[256] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[257] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020.

[258] Pavel Korshunov and Touradj Ebrahimi. Using warping for privacy protection in video surveillance. In *DSP*, pages 1–6. IEEE, 2013.

[259] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In *CVPR*, 2021.

[260] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 449–458, 2016.

[261] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[262] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 195–231. Springer, 2018.

[263] Kaspersky Lab. Man-in-the-middle attack on video surveillance systems. https://securelist.com/does-cctv-put-the-public-at-risk-of-cyberattack/70008/, Defcon,2014. [Online; accessed 30-April-2018].

[264] A. Lacoste, T. Boquet, N. Rostamzadeh, B. Oreshki, W. Chung, and D. Krueger. Deep prior. *arXiv preprint arXiv:1712.05016*, 2017.

[265] Lubor Ladicky, Chris Russell, Pushmeet Kohli, Philip H S Torr, Geremy Heitz, Stephen Gould, Ashutosh Saxena, Daphne Koller, Jim Rodgers, David Cohen, Gal Elidan, Daphne Koller, John Lafferty, Andrew McCallum, Fernando Pereira, Xuming He, Richard S Zemel, and Miguel Á Carreira-Perpiñán. Graph cut based inference with co-occurrence statistics. *International Conference on Machine Learning*, 1(3):300–316, 2009.

[266] F. Lang, T. Plötz, and S. Roth. Robust multi-image HDR reconstruction for the modulo camera. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10496 LNCS, pages 78–89, 2017.

[267] J. Laska, M. Davenport, and R. Baraniuk. Exact signal recovery from sparsely corrupted measurements through the pursuit of justice. pages 1556–1560, 2009.

[268] Tatiana Latychevskaia. Iterative phase retrieval for digital holography. *JOSA A*, 36(12):D31–D40, 2019.

[269] J. Laurent, D. Hammond, and J. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *IEEE Trans. Inform. Theory*, 57(1):559–571, 2011.

[270] J. Laurent and C. Valerio. Time for dithering: fast and quantized random embeddings via the restricted isometry property. *CoRR*, abs/1607.00816, 2016.

[271] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[272] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3361–3368. IEEE, 2011.

[273] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[274] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, 1998.

[275] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[276] Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. Riddle: Reversible and diversified de-identification with latent encryptor. In *CVPR*, 2023.

[277] Huichen Li, Linyi Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Nonlinear gradient estimation for query efficient blackbox attack. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2021), Proceedings of Machine Learning Research. PMLR*, pages 13–15, 2021.

[278] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Qeba: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1221–1230, 2020.

[279] Jianan Li, Yunchao Wei, Xiaodan Liang, Jian Dong, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Attentive contexts for object detection. *IEEE Transactions on Multimedia*, 19(5):944–954, 2016.

[280] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *CVPR*, 2020.

[281] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 641–649, 2020.

[282] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[283] Qizhang Li, Yiwen Guo, and Hao Chen. Practical no-box adversarial attacks against dnns. *Advances in Neural Information Processing Systems*, 33:12849–12860, 2020.

[284] Shasha Li, Abhishek Aich, Shitong Zhu, M. Salman Asif, Chengyu Song, Amit Roy-Chowdhury, and Srikanth Krishnamurthy. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Advances in Neural Information Processing Systems*, 34:2085–2096, 2021.

[285] Shasha Li, Karim Khalil, Rameswar Panda, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and Ananthram Swami. Measurement-driven security analysis of imperceptible impersonation attacks. *arXiv preprint arXiv:2008.11772*, 2020.

[286] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458*, 2018.

[287] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and Ananthram Swami. Stealthy adversarial perturbations against real-time video classification systems. In *NDSS*, 2019.

[288] Shasha Li, Shitong Zhu, Sudipta Paul, Amit Roy-Chowdhury, Chengyu Song, Srikanth Krishnamurthy, Ananthram Swami, and Kevin S Chan. Connecting the dots: Detecting adversarial perturbations using context inconsistency. In *European Conference on Computer Vision*, pages 396–413. Springer, 2020.

[289] Tao Li and Chris Clifton. Differentially private imaging via latent space manipulation. In *SP*, 2021.

[290] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *CVPR workshop*, 2019.

[291] X. Li. Compressed sensing and matrix completion with constant proportion of corruptions. 37(1):73–99, 2013.

[292] X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM J. on Math. Analysis*, 45(5):3019–3033, 2013.

[293] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5772, 2017.

[294] Yuezun Li, Daniel Tian, Ming-Ching Chang, Xiao Bian, and Siwei Lyu. Robust adversarial perturbation on deep proposal-based models. *arXiv preprint arXiv:1809.05962*, page 231, 2018.

[295] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and Xiaofeng Wang. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, 2018.

[296] D. Liang, J. Cheng, Z. Ke, and L. Ying. Deep mri reconstruction: Unrolled optimization algorithms meet neural networks. *arXiv preprint arXiv:1907.11711*, 2019.

[297] Siyuan Liang, Longkang Li, Yanbo Fan, Xiaojun Jia, Jingzhi Li, Baoyuan Wu, and Xiaochun Cao. A large-scale multiple-objective method for black-box attack against object detection. In *European Conference on Computer Vision*, pages 619–636. Springer, 2022.

[298] Siyuan Liang, Baoyuan Wu, Yanbo Fan, Xingxing Wei, and Xiaochun Cao. Parallel rectangle flip attack: A query-based black-box attack against object detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7677–7687. IEEE, 2021.

[299] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018.

[300] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[301] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[302] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[303] Bo Liu, Ming Ding, Hanyu Xue, Tianqing Zhu, Dayong Ye, Li Song, and Wanlei Zhou. Dp-image: differential privacy for image data in feature space. *arXiv:2103.07073*, 2021.

[304] Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, and Nenghai Yu. Detection based defense against adversarial examples from the steganalysis point of view. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4825–4834, 2019.

[305] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[306] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.

[307] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *AAAI Workshop on Artificial Intelligence Safety*, 2019.

[308] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385, 2018.

[309] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

[310] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[311] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6985–6994, 2018.

[312] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6985–6994, 2018.

[313] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[314] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[315] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

[316] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868. IEEE, 2019.

[317] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[318] Rushi Longadge and Snehalata Dongre. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*, 2013.

[319] Nicholas A. Lord, Romain Mueller, and Luca Bertinetto. Attacking deep networks with surrogate-based adversarial black-box methods is easy. In *International Conference on Learning Representations*, 2022.

[320] Nicholas A. Lord, Romain Mueller, and Luca Bertinetto. Gfcs. https://github.com/fiveai/GFCS, 2022. [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License].

[321] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 446–454, 2017.

[322] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017.

[323] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7363–7372, 2019.

[324] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

[325] Chen Ma, Li Chen, and Jun-Hai Yong. Simulating unknown target models for query-efficient black-box attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11835–11844, 2021.

[326] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[327] A. Maiden and J. Rodenburg. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy*, 109(10):1256–1262, 2009.

[328] Santiago Manen, Michael Gygli, Dengxin Dai, and Luc Van Gool. Pathtrack: Fast trajectory annotation with path supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 290–299, 2017.

[329] Oge Marques, Elan Barenholtz, and Vincent Charvillat. Context modeling in computer vision: techniques, implications, and applications. *Multimedia Tools and Applications*, 51(1):303–339, 2011.

[330] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *ARTR*, 42(2), 2014.

[331] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *CVPR*, 2020.

[332] Chris McCool, Tristan Perez, and Ben Upcroft. Mixtures of lightweight deep convolutional neural networks: Applied to agricultural robotics. *IEEE Robotics and Automation Letters*, 2(3):1344–1351, 2017.

[333] Michael McCoyd and David Wagner. Spoofing 2d face detection: Machines see people who aren't there. *arXiv preprint arXiv:1608.02128*, 2016.

[334] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 135–147, 2017.

[335] Thomas Mensink, Efstratios Gavves, and Cees G.M. Snoek. COSTA : Co-Occurrence Statistics for Zero-Shot Classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[336] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

[337] C. A. Metzler and G. Wetzstein. Deep S^3PR: Simultaneous source separation and phase retrieval using deep generative models. *arXiv preprint arXiv:2002.05856*, 2020.

[338] Christopher A Metzler, Felix Heide, Prasana Rangarajan, Muralidhar Madabhushi Balaji, Aparna Viswanath, Ashok Veeraraghavan, and Richard G Baraniuk. Deep-inverse correlography: towards real-time high-resolution non-line-of-sight imaging. *Optica*, 7(1):63–71, 2020.

[339] Christopher A Metzler, Philip Schniter, Ashok Veeraraghavan, and Richard G Baraniuk. prdeep: Robust phase retrieval with a flexible deep network. pages 3501–3510, 2018.

[340] Jianwei Miao, Tetsuya Ishikawa, Qun Shen, and Thomas Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annu. Rev. Phys. Chem.*, 59:387–410, 2008.

[341] R. Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

[342] David J Miller and Hasan Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. *NeurIPS*, 9, 1996.

[343] DL Misell. A method for the solution of the phase problem in electron microscopy. *Journal of Physics D: Applied Physics*, 6(1):L6, 1973.

[344] Michinari Momma, Chaosheng Dong, and Jia Liu. A multi-objective/multi-task learning framework induced by pareto stationarity. In *ICML*. PMLR, 2022.

[345] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *arXiv preprint arXiv:1912.10557*, 2019.

[346] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773. IEEE, 2017.

[347] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[348] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. *arXiv preprint arXiv:1707.05572*, 2017.

[349] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R Venkatesh Babu. Nag: Network for adversary generation. *arXiv preprint arXiv:1712.03390*, 2017.

[350] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[351] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam Gyu Cho, Seong Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[352] Ali Mousavi and Richard G Baraniuk. Learning to invert: Signal recovery via deep convolutional networks. *arXiv preprint arXiv:1701.03891*, 2017.

[353] Ali Mousavi, Ankit B Patel, and Richard G Baraniuk. A deep learning approach to structured signal recovery. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 1336–1343. IEEE, 2015.

[354] Baurzhan Muminov and Luat T Vuong. Small-brain neural networks rapidly solve inverse problems with vortex fourier encoders. *arXiv preprint arXiv:2005.07682*, 2020.

[355] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[356] Nina Narodytska and Shiva Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1318. IEEE, 2017.

[357] Seyedehsara Nayer, Praneeth Narayanamurthy, and Namrata Vaswani. Phaseless pca: Low-rank matrix recovery from column-wise phaseless measurements. In *International Conference on Machine Learning*, pages 4762–4770, 2019.

[358] D. Needell and J. Tropp. Cosamp: iterative signal recovery from incomplete and inaccurate samples. *Comm. of the ACM*, 53(12):93–100, 2010.

[359] ZD Net. Surveillance cameras sold on Amazon infected with malware. https://www.zdnet.com/article/amazon-surveillance-cameras-infected-with-malware/, ZD Net,2016. [Online; accessed 30-April-2018].

[360] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. pages 2796–2804, 2013.

[361] Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *TOCHI*, 13(1), 2006.

[362] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *TKDE*, 17(2), 2005.

[363] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019.

[364] David D Nolte. *Optical interferometry for biology and medicine*, volume 1. Springer Science & Business Media, 2011.

[365] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011.

[366] H. Ohlsson, A. Yang, R. Dong, and S. Sastry. Cprl–an extension of compressive sensing to the phase retrieval problem. pages 1367–1375, 2012.

[367] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–253. IEEE, 2003.

[368] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9), 2015.

[369] Jiangmiao Pang, Kai Chen, Qi Li, Zhihai Xu, Huajun Feng, Jianping Shi, Wanli Ouyang, and Dahua Lin. Towards balanced learning for instance recognition. *International Journal of Computer Vision*, 129(5):1376–1393, 2021.

[370] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2019.

[371] Nicolas Papernot, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Fartash Faghri, Alexander Matyasko, Karen Hambardzumyan, Yi-Lin Juang, Alexey Kurakin, Ryan Sheatsley, et al. cleverhans v2. 0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 2016.

[372] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[373] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[374] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[375] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[376] IS Park, RJC Middleton, Charles R Coggrave, Pablo D Ruiz, and Jeremy M Coupland. Characterization of the reference wave in a compact digital holographic camera. *Applied optics*, 57(1):A235–A241, 2018.

[377] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library.

[378] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[379] Sujoy Paul, Jawadul H Bappy, and Amit K Roy-Chowdhury. Non-uniform subset selection for active learning in structured data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 830–839. IEEE, 2017.

[380] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv:2005.05535*, 2020.

[381] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018.

[382] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8571–8580, 2018.

[383] Hugo Proença. The uu-net: Reversible face de-identification for visual surveillance video footage. *TCSVT*, 32(2), 2021.

[384] Yuying Qiu, Zhiyi Niu, Biao Song, Tinghuai Ma, Abdullah Al-Dhelaan, and Mohammed Al-Dhelaan. A novel generative model for face privacy protection in video surveillance with utility maintenance. *Applied Sciences*, 12(14):6962, 2022.

[385] Lawrence Rabiner. Fundamentals of speech recognition. *Fundamentals of speech recognition*, 1993.

[386] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. 2016.

[387] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020.

[388] Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6528–6537, 2019.

[389] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[390] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R Venkatesh Babu. Nag: Network for adversary generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 742–751, 2018.

[391] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[392] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[393] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[394] Hans Reichenbach. *Philosophic foundations of quantum mechanics*. Courier Corporation, 1998.

[395] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.

[396] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[397] J. Rhee and Y. Joo. Wide dynamic range cmos image sensor with pixel level adc. *Electron. Lett.*, 39:360–361, 2010.

[398] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2017.

[399] Yair Rivenson, Yibo Zhang, Harun Günaydın, Da Teng, and Aydogan Ozcan. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light: Science & Applications*, 7(2):17141–17141, 2018.

[400] John M Rodenburg. Ptychography and related diffractive imaging methods. *Advances in imaging and electron physics*, 150:87–184, 2008.

[401] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.

[402] Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial Patch. 2017.

[403] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[404] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *CVPR workshop*, 2018.

[405] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[406] Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, and Hamed Pirsiavash. Role of spatial context in adversarial robustness for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 784–785, 2020.

[407] Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, and Hamed Pirsiavash. Role of spatial context in adversarial robustness for object detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[408] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

[409] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

[410] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[411] K. Sasagawa, T. Yamaguchi, M. Haruta, Y. Sunaga, H. Takehara, H. Takehara, T. Noda, T. Tokuda, and J. Ohta. An implantable cmos image sensor with self-reset pixels for functional brain imaging. *IEEE Trans. on Electron Devices*, 63(1):215–222, 2016.

[412] Andrey V Savchenko. Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets. In *CVPR*, 2022.

[413] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[414] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM, 2007.

[415] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *NeurIPS*, 31, 2018.

[416] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *ICEET*, 2021.

[417] Parikshit Shah and Venkat Chandrasekaran. Iterative projections for signal identification on manifolds: Global recovery guarantees. pages 760–767. IEEE, 2011.

[418] V. Shah and C. Hegde. Signal reconstruction from modulo observations (extended version of this paper). available at: http://virajshah.me/,http://home.engineering.iastate.edu/~chinmay/, 2018.

[419] V. Shah and C. Hegde. Solving Linear Inverse Problems Using GAN Priors: An Algorithm with Provable Guarantees. 2018.

[420] V. Shah, M. Soltani, and C. Hegde. Reconstruction from periodic nonlinearities, with applications to hdr imaging. pages 863–867. IEEE, 2017.

[421] F. Shamshad and A. Ahmed. Robust compressive phase retrieval via deep generative priors. *arXiv preprint arXiv:1808.05854*, 2018.

[422] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540. ACM, 2016.

[423] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. *arXiv preprint arXiv:1801.00349*, 2017.

[424] Y. Shechtman, Y. Eldar, O. Cohen, H. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: a contemporary overview. 32(3):87–109, 2015.

[425] Milen Shishkov, Brett Eugene Bouma, and Guillermo J Tearney. System and method for optical coherence imaging, April 29 2008. US Patent 7,366,376.

[426] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020.

[427] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[428] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[429] M. Soltani and C. Hegde. Fast algorithms for demixing sparse signals from nonlinear observations. *arXiv preprint arXiv:1608.01234*, 2016.

[430] M. Soltani and C. Hegde. Stable recovery of sparse vectors from random sinusoidal feature maps. *arXiv preprint arXiv:1701.06607*, pages 6384–6388, 2017.

[431] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

[432] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[433] Thomas M Strat and Martin A Fischler. Context-based vision: recognizing objects using information from both 2 d and 3 d imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1050–1065, 1991.

[434] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *arXiv preprint arXiv:1801.04264*, 2018.

[435] Jian Sun, Huibin Li, and Zongben Xu. Deep ADMM-Net for compressive sensing MRI. pages 10–18, 2016.

[436] Fnu Suya, Jianfeng Chi, David Evans, and Yuan Tian. Hybrid batch attacks: Finding black-box adversarial examples with limited queries. *USENIX Security Symposium*, 2019.

[437] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[438] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[439] Kim Taehoon. A tensorflow implementation of "deep convolutional generative adversarial networks", 2017.

[440] Tatsuki Tahara, Xiangyu Quan, Reo Otani, Yasuhiro Takaki, and Osamu Matoba. Digital holography and its multidimensional imaging applications: a review. *Microscopy*, 67(2):55–67, 2018.

[441] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

[442] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[443] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019.

[444] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *Advances in Neural Information Processing Systems*, pages 7717–7728, 2018.

[445] Yusuke Tashiro, Yang Song, and Stefano Ermon. Diversity can be transferred: Output diversification for white-and black-box attacks. In *Advances in Neural Information Processing Systems*, volume 33, pages 4536–4548, 2020.

[446] Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *European conference on computer vision*, pages 140–153. Springer, 2010.

[447] C3D Tensorflow. C3D Implementation. https://github.com/hx173149/C3D-tensorflow.git, 2016. [Online; accessed 30-April-2018].

[448] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *CVPR*, 2020.

[449] Lei Tian, Xiao Li, Kannan Ramchandran, and Laura Waller. Multiplexed coded illumination for fourier ptychography with an led array microscope. *Biomedical optics express*, 5(7):2376–2389, 2014.

[450] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[451] T. Tirer and R. Giryes. Image restoration by iterative denoising and backward projections. 28(3):1220–1234, 2019.

[452] Malte Tölle, Ullrich Köthe, Florian André, Benjamin Meder, and Sandy Engelhardt. Content-aware differential privacy with conditional invertible neural networks. In *DeCaF*. Springer, 2022.

[453] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003.

[454] Antonio Torralba, Kevin P Murphy, and William Freeman. Contextual models for object detection using boosted random fields. In *Advances in Neural Information Processing Systems*, volume 17, pages 1401—-1408, 2005.

[455] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[456] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4489–4497. IEEE, 2015.

[457] Vikas Tripathi, Ankush Mittal, Durgaprasad Gangodkar, and Vishnu Kanth. Real time security framework for detecting abnormal events at atm installations. *Journal of Real-Time Image Processing*, pages 1–11, 2016.

[458] Doris Y Tsao and Margaret S Livingstone. Mechanisms of face perception. *Annual review of neuroscience*, 31:411, 2008.

[459] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019.

[460] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[461] E. van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

[462] E. van den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. on Sci. Computing*, 31(2):890–912, 2008.

[463] Gul Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[464] D. V. Veen, A. Jalal, M. Soltanolkotabi, E. Price, S. Vishwanath, and A. G. Dimakis. Compressed sensing with deep image prior and learned regularization. *arXiv preprint arXiv:1806.06438*, 2018.

[465] S. Venkatakrishnan, C. Bouman, and B. Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conf. on Signal and Inf. Processing*, pages 945–948. IEEE, 2013.

[466] R. Vershynin. Introduction to the non-ptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[467] Umboo Computer Vision. Case studies. https://news.umbocv.com/case-studies/home, 2016. [Online; accessed 30-April-2018].

[468] Umboo Computer Vision. Case study: Elementary scholl in taiwai. https://news.umbocv.com/case-study-taiwan-elementary-school-13fa14cdb167, 2017. [Online; accessed 30-April-2018].

[469] Umboo Computer Vision. Umbo Customer Case Study NCHU. https://news.umbocv.com/umbo-customer-case-study-nchu-687356292f43, 2017. [Online; accessed 30-April-2018].

[470] Umboo Computer Vision. Umbo's smart city featured on cbs sacramento. https://news.umbocv.com/umbos-smart-city-featured-on-cbs-sacramento-26f839415c51, 2017. [Online; accessed 30-April-2018].

[471] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed.*, 2017.

[472] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.

[473] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[474] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12645–12654, 2020.

[475] G. Wang and G. Giannakis. Solving random systems of quadratic equations via truncated generalized gradient flow. pages 568–576, 2016.

[476] G. Wang, L. Zhang, G. B. Giannakis, M. Akcakaya, and J. Chen. Sparse phase retrieval via truncated amplitude flow. 66:479–491, 2018.

[477] Gang Wang, Georgios Giannakis, Yousef Saad, and Jie Chen. Solving most systems of random quadratic equations. In *Advances in Neural Information Processing Systems*, pages 1867–1877, 2017.

[478] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[479] Jie Wang, Qinhua Gao, Xiaorui Ma, Yunong Zhao, and Yuguang Fang. Learning to sense: Deep learning for wireless sensing with less training efforts. *IEEE Wireless Communications*, 2020.

[480] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 43(10), 2021.

[481] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Proximal deep structured models. In *Advances in Neural Information Processing Systems*, pages 865–873, 2016.

[482] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

[483] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[484] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021.

[485] Yajie Wang, Yu an Tan, Wenjiao Zhang, Yuhang Zhao, and Xiaohui Kuang. An adversarial attack on DNN-based black-box object detectors. *Journal of Network and Computer Applications*, 161:102634, 2020.

[486] Yue Wang, Cheng Si, and Xintao Wu. Regression model fitting under differential privacy and model inversion attack. In *IJCAI*, 2015.

[487] James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[488] Ke Wei. Solving systems of phaseless equations via kaczmarz methods: A proof of concept study. *Inverse Problems*, 31(12):125008, 2015.

[489] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018.

[490] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 954–960, 2019.

[491] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8973–8980, 2019.

[492] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12338–12345, 2020.

[493] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *arXiv preprint arXiv:1511.04136*, 2015.

[494] Yunqian Wen, Bo Liu, Ming Ding, Rong Xie, and Li Song. Identitydp: Differential private identification protection for face images. *Neurocomputing*, 501, 2022.

[495] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: multi-head cross attention network for facial expression recognition. *arXiv:2109.07270*, 2021.

[496] Mika Westerlund. The emergence of deepfake technology: A review. *TIM Review*, 9(11), 2019.

[497] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *International Conference on Learning Representations*, 2020.

[498] S. Wu, A. Dimakis, F. Sanghavi, S.and Yu, D. Holtmann-Rice, D. Storcheus, A. Rostamizadeh, and S. Kumar. Learning a compressed sensing measurement matrix via gradient unrolling. pages 6828–6839, 2019.

[499] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018.

[500] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.

[501] Chaowei Xiao, Ruizhi Deng, Bo Li, Taesung Lee, Benjamin Edwards, Jinfeng Yi, Dawn Song, Mingyan Liu, and Ian Molloy. Advit: Adversarial frames identifier based on temporal consistency in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3968–3977, 2019.

[502] Chaowei Xiao, Ruizhi Deng, Bo Li, Fisher Yu, Mingyan Liu, and Dawn Song. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–234, 2018.

[503] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[504] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[505] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

[506] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1369–1378, 2017.

[507] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019.

[508] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.

[509] Jiangjian Xie, Jun Yang, Changqing Ding, and Wenbin Li. High accuracy individual identification model of crested ibis (nipponia nippon) based on autoencoder with self-attention. *IEEE Access*, 8:41062–41070, 2020.

[510] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[511] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *CVPR*, 2022.

[512] Dan Xu, Rui Song, Xinyu Wu, Nannan Li, Wei Feng, and Huihuan Qian. Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. *Neurocomputing*, 143:144–152, 2014.

[513] Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2014.

[514] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Annual Network and Distributed System Security Symposium*, 2017.

[515] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *CVPR*, 2022.

[516] Zongben Xu and Jian Sun. Image inpainting by patch propagation using patch sparsity. *IEEE transactions on image processing*, 19(5):1153–1165, 2010.

[517] T. Yamaguchi, H. Takehara, Y. Sunaga, M. Haruta, M. Motoyama, Y. Ohta, T. Noda, K. Sasagawa, T. Tokuda, and J. Ohta. Implantable self-reset cmos image sensor and its application to hemodynamic response detection in living mouse brain. *Japanese J. of Appl. Physics*, 55(4S):04EM02, 2016.

[518] Jiancheng Yang, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, and Chenglong Zhao. Learning black-box attackers with transferable priors and query feedback. *Advances in Neural Information Processing Systems*, 33:12288–12299, 2020.

[519] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *CVPR*, 2019.

[520] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing mri. In *Advances in neural information processing systems*, pages 10–18, 2016.

[521] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019.

[522] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 702–709, 2012.

[523] L. Yeh, J. Dong, J.and Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller. Experimental robustness of fourier ptychography phase retrieval algorithms. *Optics express*, 23(26):33214–33240, 2015.

[524] Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016.

[525] Mingjun Yin, Shasha Li, Zikui Cai, Chengyu Song, M. Salman Asif, Amit K Roy-Chowdhury, and Srikanth V Krishnamurthy. Exploiting multi-object relationships for detecting adversarial attacks in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7858–7867, 2021.

[526] Jinao Yu, Hanyu Xue, Bo Liu, Yu Wang, Shibing Zhu, and Ming Ding. Gan-based differential private image privacy protection framework for the internet of multimedia things. *Sensors*, 21(1), 2020.

[527] Z. Yuan and H. Wang. Phase retrieval with background information. *Inverse Problems*, 35(5):054003, may 2019.

[528] Zheng Yuan, Jie Zhang, Yunpei Jia, Chuanqi Tan, Tao Xue, and Shiguang Shan. Meta gradient adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7748–7757, 2021.

[529] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[530] Guangsheng Zhang, Bo Liu, Tianqing Zhu, Andi Zhou, and Wanlei Zhou. Visual privacy attacks and defenses in deep learning: a survey. *Artificial Intelligence Review*, 55(6), 2022.

[531] H. Zhang and Y. Liang. Reshaped wirtinger flow for solving quadratic system of equations. pages 2622–2630, 2016.

[532] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[533] Hantao Zhang, Wengang Zhou, and Houqiang Li. Contextual Adversarial Attacks for Object Detection. *International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

[534] Hu Zhang, Linchao Zhu, Yi Zhu, and Yi Yang. Motion-excited sampler: Video adversarial attack with sparked prior. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[535] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5678–5686, 2017.

[536] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[537] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep cnn denoiser prior for image restoration. pages 3929–3938, 2017.

[538] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020.

[539] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. FreeAnchor: Learning to match anchors for visual object detection. In *Neural Information Processing Systems*, pages 147–155, 2019.

[540] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *ECCV*, 2020.

[541] H. Zhao, B. Shi, C. Fernandez-Cull, S. Yeung, and R. Raskar. Unbounded high dynamic range photography using a modulo camera. In *Intl. Conf. on Comp. Photography (ICCP)*, 2015.

[542] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[543] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018.

[544] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[545] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1989–2004, 2019.

[546] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[547] Jizhe Zhou and Chi-Man Pun. Personal privacy protection via irrelevant faces tracking and pixelation in video live streaming. *TIFS*, 16, 2020.

[548] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.

[549] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv:2005.10353*, 2020.

[550] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. pages 597–613. Springer, 2016.

[551] Shitong Zhu, Zhongjie Wang, Xun Chen, Shasha Li, Umar Iqbal, Zhiyun Qian, Kevin S Chan, Srikanth V Krishnamurthy, and Zubair Shafiq. A4: Evading learning-based adblockers. *arXiv preprint arXiv:2001.10999*, 2020.

[552] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

[553] Yingying Zhu, Nandita M Nayak, and Amit K Roy-Chowdhury. Context-aware activity recognition and anomaly detection in video. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):91–101, 2012.

[554] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *CVPR*, 2021.

[555] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 593–602, 2019.