UNIVERSITY OF CALIFORNIA
RIVERSIDE

# Robustness of Vision-Language Systems to Intentional and Incidental Adversity

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

Doctor of Philosophy

in

Computer Science

by

Ghazal Mazaheri

December 2022

Dissertation Committee:

Dr. Amit K. Roy-Chowdhury, Co-Chairperson
Dr. Evangelos Papalexakis , Co-Chairperson
Dr. Jiasi Chen
Dr. Tamar Shinar

The Dissertation of Ghazal Mazaheri is approved:

---

---

---

Committee Co-Chairperson

---

Committee Co-Chairperson

University of California, Riverside

## Acknowledgments

This dissertation would not have been possible without the support and help of many people in my life. First and foremost, I would like to thank my advisors: Prof. Amit Roy-Chowdhury and Prof. Evangelos Papalexakis. Prof. Amit Roy-Chowdhury gave me a second chance, without which I would never have made it here. His bottomless well of support and encouragement has been an ever-present source of strength for me, for which I will be forever grateful. I express my deepest gratitude to him for his inspiration and guidance in my research, writing and presentations. For his understanding and patience through my highs and my lows. I would like to thank Prof. Evangelos Papalexakis for his invaluable and to the point comments. Our constructive discussions added a new point of view to research solutions and boosted the quality of my research. I believe many students at UCR are attracted to data science because of his exciting, five-star data mining course and I am not an exception. This dissertation is a culmination of my advisors' constant support, motivation and training throughout my PhD. I feel extremely fortunate and privileged to be part of UCR and work under their supervision.

I would also like to express my gratitude to my dissertation committee members, Dr. Jiasi Chen, and Dr. Tamar Shinar for giving me valuable feedback and support. I would like to thank Dr. Subarna Tripathi for the collaboration and guidance during my research. Sudipta Paul has simultaneously been a friend, and co-author and I am grateful for him helping me choosing research directions and ideas to pursue. I would like to express my sincere thanks to all my lab-mates over the years. You guys were always there to lend an ear to my frustrations or ideas. You're a constant source of inspiration.

I remain ever grateful to my mom, Roya, who is the key instigator in fueling my desire to pursue graduate studies. None of this would have been possible without her unconditional love, understanding, co-operation and constant emotional support throughout my PhD. At my lowest of lows, she has been the reason to see the light at end of the tunnel and I miss her hug more than anything in this world. She has been my best friend and a great role model.

To all brave Iranian women.

# ABSTRACT OF THE DISSERTATION

Robustness of Vision-Language Systems to Intentional and Incidental Adversity

by

Ghazal Mazaheri

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, December 2022
Dr. Amit K. Roy-Chowdhury, Co-Chairperson
Dr. Evangelos Papalexakis , Co-Chairperson

The availability of large data collections makes deep neural networks perform impressively on a wide range of vision-language tasks. However, existence of corrupted samples when working with massive, hard-to-manage amount of data is inevitable. Data corruption refers to errors in data that occur during either an intentional process to alter data samples with the purpose of hiding information or a incidental data preparing step, including data annotation, which introduces unintended changes to the original data. Robustness of vision-language systems to intentional and incidental adversity is the goal of this dissertation.

Face manipulation is one of the challenging intentional adversity techniques that distort the truth. The primary threat of face manipulation stems from people becoming convinced that something fictional really occurred. The dissertation starts with study of facial manipulation detectors leveraging upon facial expression recognition systems. Concerns regarding the wide-spread use of corrupted images and videos in social media necessitate precise detection of such fraud. Multi-task learning can leverage prominent features learned by facial expression recognition system to benefit the training of conventional manipulation

detection systems. Such an approach achieves impressive performance in facial expression manipulation detection, while exhibiting robustness to other common facial corruption mechanisms, such as identity manipulation.

Creation of large scale datasets, needed for tasks like image-text retrieval, is challenging. As a solution, recent works exploit vision-language pre-training to overcome the errors in the training data. Vision-language pre-training has advanced the performance for many vision-language tasks like image-text retrieval. However, performance improvement has been largely achieved by scaling up the dataset with noisy image-text pairs collected from the web, which is a suboptimal source of supervision. Web-crawled captions do not accurately describe the visual content of the images, making them noisy signals. Instead, the caption descriptions can be generated by off-the-shelf automatic image captioners.

In this dissertation, we focus on learning image-text retrieval models utilizing image-text pairs, where the text is generated by an automatic captioning process without any requirement of tedious annotation effort. We propose a novel robust image-text retrieval model which utilizes data-driven curriculum learning to down-weight the noisy portion of the dataset generated by image captioners. Finally, we investigate fine-grained image-text retrieval which decomposes image-text matching into global-to-local level matching using pre-trained object detectors. We utilize an alignment model which bridges the gap between corresponding embeddings in different modalities to identify the noise in the caption. It aims to learn image-text multimodal representations that capture the appropriate fine-grained alignment between vision and language.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The most recent developments in Machine Learning (ML), especially in Deep Learning, evolved from the concept of a single-layer neural network, the perceptron [102], to the Multi-Layer Perceptron (MLP) [102] and the current intricate multi-layer Deep Neural Networks (DNNs) [123, 48], with the objective of approaching and even exceeding human decision making capabilities for a certain set of tasks. Due to their effectiveness at handling large amounts of data, learning (and sometimes re-learning [7]) input characteristics, and demonstrating high accuracy during inference of unseen inputs, ML systems have proliferated to numerous real-world applications. These include object detection [37], face recognition [46], manipulation detection [134], vision-language tasks like image-text retrievals [69] and even safety-critical applications like autonomous driving [14], and health-care [8], where errors may lead to catastrophic results.

Poisoning the data by inserting malicious and crafted samples is one way in which the security and integrity can be can breached. Despite high inference accuracy in practical

applications, ML systems especialy DNNs are highly vulnerable to to maliciously corrupted examples in datasets. Polluting inputs with incorrectly-labeled samples even in the absence of an explicit attacker during training and inference can compromise the reliability of an ML system. Therefore, some of data corruption techniques are intentionally designed to mislead viewers into thinking something fake is real and threat people's security. Others occur unintentionally during data collection to generate large-scale datasets. Approaches to defend ML systems against these concerns remains a challenging problem, which we will discuss in this thesis.

A common term associated with the performance of DNNs is robustness [113]. Robustness is the DNN property that determines the integrity of the network under varying operating conditions, and the accuracy of DNN outputs in the presence/absence of input or network alterations [113]. Correctly quantifying the robustness of machine learning models is a central aspect in judging their suitability for specific tasks, and thus, ultimately, for generating trust in the models. This can be divided into two sub-properties: security and reliability. The DNN is said to be secure against an attack if the attacker cannot steal information, modify the network parameters, or render an incorrect input to the DNN (e.g., using an adversarial attack) [98]. In case of reliability, there is no explicit attacker. The network is said to be reliable if it does not display any changes to its output, parameters, or behavior if one or more of the input independent variables (features) or assumptions are drastically changed. In this thesis, our focus is on reliability of vision-language Systems to intentional and incidental adversity.

Face manipulation is one of the challenging intentional data corruption techniques that distort the truth. The primary threat of face manipulation stems from people becoming convinced that something fictional really occurred. The dissertation starts with study of facial manipulation detectors leveraging upon facial expression recognition systems. Concerns regarding the wide-spread use of forged images and videos in social media necessitate precise detection of such fraud. Facial manipulations can be created by identity swap (DeepFake) or expression swap. Most of recent face forensic works focus on deepfake detection, detecting whether identity of faces are real or artificially generated [104, 6, 27, 67]. The very popular term "DeepFake" is referred to a deep learning based technique able to create fake videos by swapping the face of a person by the face of another person.

Contrary to the identity swap, which can easily be detected with novel deepfake detection methods, expression swap detection has not yet been addressed extensively. The importance of facial expressions in inter-person communication is known. Consequently, it is important to develop methods that can detect and localize manipulations in facial expressions. To this end, in Chapter 2, we present a novel framework to exploit the underlying feature representations of facial expressions learned from expression recognition models to identify the manipulated features. Using discriminative feature maps extracted from a facial expression recognition framework, our manipulation detector is able to localize the manipulated regions of input images and videos. Such approach achieves impressive performance in facial expression manipulation detection, while exhibiting robustness to other common facial corruptions which degrade other models' performance.

In addition to robust systems toward intentional corrupted data, it is imperative to design machine learning systems which are robust to incidental noise in the training data. Creation of large scale data sets, needed for the tasks like image-text retrieval, pose a challenge to tackle incidental noise. Since well-annotated datasets can be prohibitively expensive and time-consuming to collect, recent work has explored the use of larger but noisy datasets that can be more easily obtained [52, 112].

There is a recent growing interest in pre-training generic models to solve a variety of vision-language problems, such as visual question answering, image-text retrieval and image captioning etc. vision-language pre-training aims to improve performance of downstream vision and language tasks by pre-training the model on large-scale image-text pairs. For example, there are different architectures (e.g. two-stream models [56, 81, 80] vs. single-stream models [21, 63]), backbones (e.g. ConvNets [50] vs. Transformers [57]) etc. All these works aim to exploit the large-scale aligned image-text corpora to pre-train a powerful multi-modal model, which is then adapted to various downstream vision-language tasks. Despite the use of simple rule-based filters, noise is still prevalent in large-scale texts crawled from web. However, merely scaling up the dataset may not be possible in many applications, like agriculture and environmental monitoring or where we deal with private data (banking, healthcare, etc.) where large volumes of (even poorly) labeled data may not be available. In this thesis, we focus on learning image-text retrieval models utilizing image-text pairs, where the text is generated by an automatic captioning process without any requirement of tedious annotation effort. However, such generated descriptions are likely to be noisy (e.g.,wrong detection of objects, misinterpretation of the image concept) due to the use of

automated captioning process. In Chaper 3 and 4, we propose methods to learn directly from noisy captions without any pre-processing step.

In Chapter 3, we propose a robust image-text retrieval learning approach. To learn a retrieval model robust to noisy captions, it is imperative that the model is able to emphasize the clean part of the dataset and learn from it, while avoiding the noisy part. To this end, inspired by the success of Curriculum Learning (CL) to handle label noise [55], we propose a data-driven curriculum learning method using external knowledge. It learns to weigh the samples in a way that the model can focus on the samples whose captions have a high chance of being correct. We use a region encoder and a label encoder trained on an external source of data with correct annotation to obtain an approximation of how likely an image-text pair is correct. This approximation is used to guide the data-driven curriculum learning process which can generate sample weights. Such sample weights, learned during training, compel the framework to learn from clean examples. In this thesis, we specifically focus on the scenario where the textual description is noisy due to automatic captioning process similar to alt text generation.

In Chapter 4, our proposed method for improving the robustness of image-text retrieval system utilizes an alignment model which bridges the gap between corresponding embeddings in different modalities to identify the noise in the caption. In general, most existing cross-modal matching methods embed different modalities into a common space wherein the similarity of positive cross-modal pairs is maximized and that of the negative ones is minimized. Although these methods have achieved promising results, their success depends on an implicit data assumption, i.e., the training data are correctly aligned across

modalities. For example, in the vision-language tasks, the text needs to accurately describe the image content, and vice versa. However, noise in the captions which is inevitable when they are generated using captioning model, leads to difficulty in bridging the heterogeneity gap [137, 77] between image and text modality.

The challenge of robust image-text retrieval systems is that simple joint embeddings are insufficient to represent complicated visual and textual details, such as scenes, objects, their compositions, especially with presence of noise in the caption. To tackle this issue, we gain motivation from fine-grained image-text retrievals which decompose image-text matching into global-to-local level matching using pre-trained object detectors. We equip our image-text retrieval system with alignment model which associates automatically detected tags/object labels with the corresponding image regions to identify the noise in the caption. Our approach not only takes comprehensive and fine-grained cross-modal interactions into account, but also properly handles negative pairs and irrelevant information. Extensive experiments are conducted on well-known image-text datasets that demonstrate the effectiveness of the proposed approaches compared to multiple state-of-the-art methods for both image-to-text and text-to-image retrieval.

**Organization of the Thesis.** The rest of the thesis is organized as follows. In Chapter 2, we address the problem of facial expression manipulation detection. We present a novel framework to exploit the underlying feature representations of facial expressions learned from expression recognition models to identify the manipulated features. Such approach achieves impressive performance in facial expression manipulation detection, while exhibiting robustness to other common facial corruptions like identity manipulation. In Chapters

3 and 4, we propose robust image-text retrieval learning approaches. In Chapter 3, our proposed framework utilizes data-driven curriculum learning to up-weight the clean samples while discounting the noisy portion of the training data. Such sample weights, learned during training, compel the framework to learn from clean examples. In Chapter 4, we equip our image-text retrieval system with an alignment model which bridges the gap between corresponding embeddings in different modalities to identify the noise in the caption. It aims to learn image-text multimodal representations that capture the appropriate fine-grained alignment between vision and language. We conclude the thesis in Chapter 5 by providing some future directions related to the problem of robust image-text retrieval systems.

# Chapter 2

# Detection and Localization of Facial Expression Manipulations

Concerns regarding the wide-spread use of forged images and videos in social media necessitate precise detection of such fraud. Facial manipulations can be created by Identity swap (DeepFake) or Expression swap. Contrary to the identity swap, which can easily be detected with novel deepfake detection methods, expression swap detection has not yet been addressed extensively. The importance of facial expressions in inter-person communication is known. Consequently, it is important to develop methods that can detect and localize manipulations in facial expressions.

To this end, we present a novel framework to exploit the underlying feature representations of facial expressions learned from expression recognition models to identify the manipulated features. Using discriminative feature maps extracted from a facial expression recognition framework, our manipulation detector is able to localize the manipulated

regions of input images and videos. On the Face2Face dataset, (abundant expression manipulation), and NeuralTextures dataset (facial expressions manipulation corresponding to the mouth regions), our method achieves higher accuracy for both classification and localization of manipulations compared to state-of-the-art methods. Furthermore, we demonstrate that our method performs at-par with the state-of-the-art methods in cases where the expression is not manipulated, but rather the identity is changed, leading to a generalized approach for facial manipulation detection.

## 2.1 Introduction

Facial expressions are critical in communicating our thoughts, ideas, emotions, and in responding to each other emotionally and physically. With effectual facial expressions, a person may convince others to believe in ideas without verbal communication. Due to the power of facial expressions in person-to-person communication, it is critical to determine if the facial expressions in an image or video are the individual's original expressions or manipulated by an external agent. Facial manipulations can be created by Identity swap (DeepFake) or Expression swap. In this thesis, *we focus on the problem of detecting facial expression manipulations while ensuring that performance does not degrade for the identity manipulation case*, thus providing a *generalized version of facial manipulation detection.*

Facial expression changes in the Face2Face dataset is a result of a facial reenactment system that transfers the expressions of a source video to a target video while maintaining the identity of the target person. Similarly, NeuralTextures [116] reenacts face motions of an input video to a target video mainly affecting the regions around the

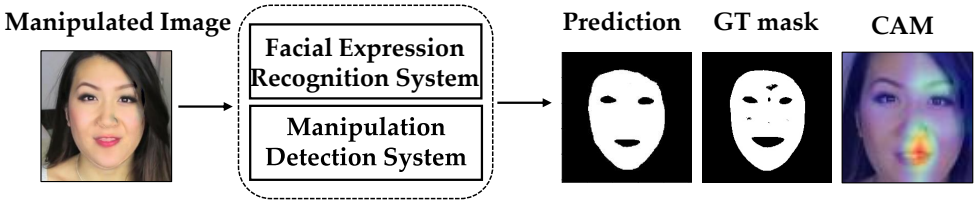



Figure 2.1: Our overall framework for detection of facial expression manipulation and localization. Manipulated mage is from Face2Face dataset [118]. Class activation map (CAM) is the visualization of feature map from facial expression recognition system.

mouth. We hypothesize that to detect such facial expression manipulations, recognition of the expression would be helpful. Based on this key idea, we design a framework called the Expression Manipulation Detection (EMD) system. Fig. 2.1 presents this key idea of using facial expression recognition to guide the manipulation detection procedure. As can be seen in the figure, the main manipulations appear in the parts of the face which constitute expression change, such as, regions around eyebrows and mouth which are critical regions for facial expressions.

In order to exploit prominent features corresponding to facial expression, we use Facial Expression Recognition (FER) systems in our face manipulation detection framework (see Sec. 2.3.3 for details). In particular, we adopt Ensemble with Shared Representations (ESR) [111] as the backbone network of FER system. Feature maps from the penultimate layer of FER systems contain important information regarding facial expressions in faces [141], which we aim to exploit in order to improve over state-of-art manipulation detection methods.

**Framework Overview** A pictorial illustration of our facial expression manipulation detection (EMD) framework is presented in Fig. 2.2. EMD utilizes a two-stream network for manipulation detection. One stream (FER) is responsible for extracting impor-

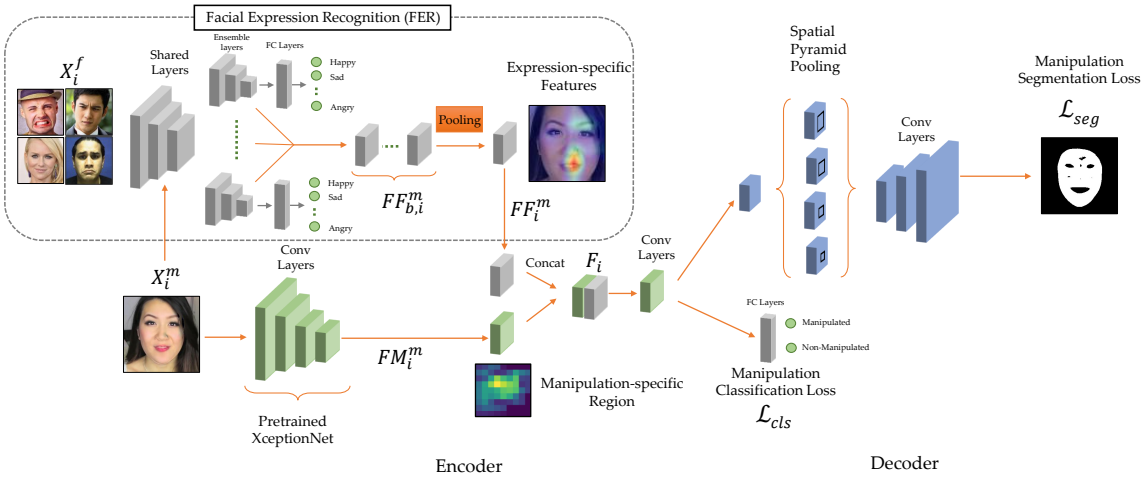



Figure 2.2: This figure represents our proposed approach for facial expression manipulation detection and localization. Extracted features from FER System ($FF_i^m$), along with the ones from manipulation detection stream ($FM_i^m$), are fed into the decoder for pixel-wise localization of the manipulated region. Notation is described in the text. The details are explained is Sec. 2.3.3 and 2.3.4

tant information for facial expressions. The feature maps from the last layer of FER stream provide information about the facial regions that encode the expression information. The other stream is an encoder-decoder architecture which is responsible for manipulation detection. The encoder projects the image to a lower dimensional space, where the features from the FER system are combined and then a decoder is used to predict the manipulated regions (if any) of the facial image. Our FER system uses ESR [111] to extract expression relevant features. The penultimate layer feature map of the FER system contains features which are discriminative to detect relevant portions of the image specific for expression. The second stream, i.e., the manipulation detection and segmentation system is an encoder-decoder architecture. The decoder takes the latent space features from the encoder and FER system combined and projects it using DeepLabv3+ [18] for manipulation localization. (see Fig. 2.2).

**Main contributions.** We propose a novel approach for facial expression manipulation detection leveraging upon a facial expression recognition system. This leads to higher performance in forgery detection where the facial expression is manipulated, as well as localizing the regions that have been manipulated. Our system achieves at par performance in the case where the identity is manipulated, ensuring generalizability of our method. Our method leads to more than 3% improvement in manipulation classification and localization over the state of the art on the Face2Face dataset [118] where the expressions are manipulated. We also show the effectiveness of our method by presenting the results on NeuralTextures dataset [116] where the facial expressions corresponding to the mouth regions have been modified. On NeuralTextures dataset, we achieve 2% higher accuracy for classification and localization. Finally, our framework achieves competitive results compared to the state-of-the-art methods on DeepFake dataset [5] and DFDC dataset [31] where identity, rather than expression, is manipulated.

## 2.2 Related Work

Multimedia forensics aims to ensure authenticity, origin, and provenance of an image or video. In recent years, there has been a variety of works in 1. forgery classification where a proposed method recognizes whether or not an image or video is manipulated, and 2. forgery localization which highlights the exact position of manipulated region [122]. We will briefly survey existing work in both the mentioned categories, as well as facial expression recognition. There is no work that specifically focuses on the problem of detection and localization of facial expression manipulations.

**Forgery Classification** In forgery classification area, there has been a variety of works in image manipulation detection [12, 22, 78, 97, 11, 127, 144] or fake faces classification in videos [104, 6, 27, 67]. The very first works aim to detect wide range of manipulations in images including object removal, copy-move, and splicing using handcrafted features that capture expected statistical or physics-based features which occur during image formation. Authors in [78] propose an approach for detecting copy-paste and composite forgery in JPEG images. Work in [22] investigates the influence of denoising on PRNU-based forgery detection. Inspired by the success of deep neural networks in different visual recognition tasks, deep learning-based approaches have been popular choices for image forgery detection. Some recent deep learning-based methods such as convolutional neural networks (CNN) have been applied to detect/classify image manipulations [12, 11, 127, 144].

Manipulation of faces in images/videos has been in the news lately. Manipulation detection in faces is challenging since exiting manipulation techniques leave almost no visual traces. In [27] authors propose to utilize an attention mechanism to process and improve the feature maps for the classification task. Four main categories of face manipulation techniques include Face2Face [118], DeepFake [5], FaceSwap [3] and NeuralTextures [117]. The most comprehensive dataset containing all four manipulation techniques (FaceForensics++) has been introduced in [104]. Generative networks play an important role in producing manipulated images and videos. Work in [136] proposes a new model to perform multiple facial attributes manipulation with one-input multi-output architecture.

To detect face manipulation in videos, some approaches utilize video temporal features as tampering of individual frames in videos causes inconsistency. The work in [106]

uses CNN as feature extractor and LSTM to capture video temporal features. Some other works use physiological signals, like eye blinking in [70] and head movements in [130], that are not well presented in the synthesized fake videos. Instead of using temporal features, authors in [104, 6, 71, 90] proposed methods which utilize extracted images from different frames of a video. Work in [128] proposes a visual speaker authentication scheme based on the deep convolutional neural network (DCNN) to defend against DeepFake attacks.

**Forgery Localization** Localizing the exact position of manipulated regions in an image or video provides critical additional information. There has been a variety of works that attempted to segment out tampered regions [126, 10, 145]. Early works [105, 36, 13] reveal the tampered regions using traditional image processing-based approaches. Researchers in [10, 75, 15, 138] exploit machine learning techniques in order to classify if a patch is manipulated or not.

Authors in [145] use object detection method proposed in [100] to identify fake objects. Unlike [145] which utilizes bounding box to coarsely localize manipulated object, [11] adopt a segmentation approach to segment out manipulated regions by classifying each pixel (manipulated/non-manipulated). Semantic segmentation approaches are suitable for fine-grained localization of tampered regions in an image. A typical semantic segmentation approach focuses on segmenting all meaningful regions (objects). However, a segmentation approach for localization of image manipulation needs to focus only on the possible tampered regions which bring additional challenges to an existing challenging problem. To localize tampered regions, [11] used an LSTM Encode-Decoder architecture.

Fake face segmentation is one of the recent challenges which has not yet been addressed extensively. Some of the proposed methods may have high performance in face manipulation detection [104] but do not address the task of segmenting the manipulated region. Multi-tasking approaches are promising in the combined classification and segmentation task. Work in [90] uses Y-shape architecture to classify manipulated videos and segment tampered faces simultaneously. In our proposed method, in addition to manipulation classification and segmentation stream, we add another stream as face expression recognition which operates jointly with the manipulation stream in order to exploit necessary information in faces. This improves the performances in both classification and segmentation.

**Facial Expression Recognition** The development of machine learning and the advent of deep learning have significantly improved the research of FER. There have been variety of works in literature which obtain high performance for facial expression recognition framework [44, 45, 68, 111, 129, 135]. The careful design of local to global feature learning with a convolution, pooling, and layered architecture produces a rich visual representation, making CNN a powerful tool for facial expression recognition. Research challenges such as the Emotion Recognition in the Wild (EmotiW) and Kaggle's Facial Expression Recognition Challenge suggest the growing interest of the community in the use of deep learning for the solution of this problem.

To have more accurate FER, the networks become deeper and deeper in order to deal with more complex classifications. Also, attention mechanisms are introduced in many networks to improve facial expression recognition. Authors in [114] proposed a facial

expression recognition network with the visual attention mechanism. Work in [111] is one of the most recent in this area showing promising results on facial expression datasets using shared and ensemble layers.

## 2.3 Methodology

In this section, we present our framework for facial expression manipulation detection and localization. We start with a formal description of the problem statement followed by the two streams of our framework - Facial Expression Recognition (FER) stream and the encoder-decoder based manipulation detection stream which receives information from FER for better detection.

### 2.3.1 Problem Statement

Consider we have a dataset of tuples $\mathcal{X}^M = \{(X_i^m, M_i^m, y_i^m)\}_{i=1}^N$, where $X_i^m \in \mathbb{R}^{H\times W\times 3}$ is a 2D image of faces, $M_i \in \mathbb{R}^{H\times W}$ is 2D binary mask of manipulated regions, and $y_i \in \{0, 1\}$ is an indicator whether $y_i^m$ is manipulated or not. Given such a dataset, our main goal is to learn a model that would be able to classify a test image to be either manipulated or not, and more importantly, localize portions of the image which are manipulated.

To identify manipulations in facial expressions, we need to focus on regions specific for expressions; thus, we utilize an auxiliary task of Facial Expression Recognition (FER). We use a dataset of tuples $\mathcal{X}_F = \{(X_i^f, y_i^f)\}_{i=1}^{N'}$, where $X_i^f \in \mathbb{R}^{H\times W\times 3}$ and $y_i^f \in \{1, \ldots, C\}$, and $C$ is the number of facial expression categories. Note that we use the superscripts $m$ and $f$ to denote data points from the manipulation set $\mathcal{X}^M$ and facial expression set $\mathcal{X}^F$.

### 2.3.2 Algorithm Overview

Our EMD system consists of two main parts including FER and encoder-decoder. We train the FER module using the dataset $\mathcal{X}^F$. To train the encoder-decoder architecture for manipulation detection, our framework takes information from the FER module. However, we pass the images in $\mathcal{X}^M$ through both the streams - an encoder to obtain features necessary to detect manipulations, and an FER module to obtain features specific to facial expressions. We combine these features in the latent space and then pass them through a decoder to spatially localize the manipulated regions.

### 2.3.3 Facial Expression Recognition

We utilize one of the state-of-the-art methods for facial expression recognition proposed in [111] as a pre-trained model for recognizing the facial expressions.

FER system presented in [111] consists of two building blocks. The base of the network (shared layers in Fig. 2.2) is an array of convolutional layers for low- and middle-level feature learning. These informative features are then shared with independent convolutional branches that constitute the ensemble (ensemble layers in Fig. 2.2). From this point, each branch can learn distinctive features while competing for a common resource - the shared layers. This competitive training emerges from the minimization of a combined loss function defined as the summation of the loss functions (cross-entropy loss) of each branch as follows:

$$\mathcal{L}_{FER} = \frac{1}{N'} \sum_{b} \sum_{i} \sum_{c} -y^f_{i,c} \log(p^f_{b,i,c}) \tag{2.1}$$

where given an image $X_i^f$, $p_{b,i,c}^f$ is a probability mass function over the $C$ facial expression categories for branch $b$. $N'$ is total number of images in the dataset.

After training the FER system, we use the pre-trained models to predict the expression category of manipulated images and extract the feature maps needed to be combined with the manipulation detection stream. The reason is that, for manipulation detection task, feature maps which highlight the discriminative image regions important for expression recognition are useful for manipulation detection of images where facial expressions are tampered such as manipulation in Face2Face dataset. Therefore, our FER architecture provides useful expression and location-aware features needed for the manipulation detection task.

In the Facial Expression Recognition system, we pass an image $X_i^m$ through the shared convolutional network $Conv$ and an ensemble of $B$ convolutional networks $\{\mathcal{E}_b\}_{b=1}^B$ to generate $B$ different feature maps of the facial expression. For the $b^{th}$ ensemble convolutional branch, the feature map corresponding to $X_i^m$ is,

$$FF_{b,i}^m = \mathcal{E}_b(Conv(X_i^m)). \tag{2.2}$$

We use a classifier $\mathcal{M}$ to infer class probabilities for the $C$ expression classes. For $b^{th}$ branch network and $i^{th}$ image $X_i^m$, we infer the class probability vector, $\boldsymbol{\rho}_{b,i} = [\rho_{b,i,1}, \ldots, \rho_{b,i,C}]$. Here, $\rho_{b,i,j} = \mathcal{M}(FF_{b,i}^m, c_j)$ indicates the detection probability of class $c_j$ for branch $b$ with input image $X_i^m$. Therefore, the detected expression class of an image from $b^{th}$ convolutional

branch is

$$c_{det}^{b} = \underset{c_j \in \{c_1,\ldots,c_C\}}{\arg\max}\ \mathcal{M}(FF_{b,i}^{m}, c_j). \tag{2.3}$$

Considering the detection of all the branches, most frequent detected class for an image is,

$$c_{freq} = mode(\{c_{det}^{1}, \ldots, c_{det}^{B}\}) \tag{2.4}$$

The feature map from the branch in FER network that results in highest detection probability for the frequently detected class $c_{freq}$ is pooled for manipulation detection task. So, the pooled feature map is

$$FF_{i}^{m} = \underset{FF_{b,i}^{m} \in \{FF_{1,i}^{m},\ldots,FF_{B,i}^{m}\}}{\arg\max}\ \mathcal{M}(FF_{b,i}^{m}, c_{freq}) \tag{2.5}$$

As will be discussed subsequently, we use the feature map $FF_{i}^{m}$ after convolutional layers as auxilliary input to the encoder-decoder for manipulation detection. As illustrated later in Fig. 2.5, we also obtain the class activation maps (CAMs) for visualization purpose following [141].

### 2.3.4 Encoder-Decoder

Encoder-decoder networks using CNN architecture have been extensively used in deep learning literature, specifically for semantic object segmentation. Following the literature, we adopt Encoder-Decoder architecture [18] known as deeplabv3+ for manipulation detection and segmentation as the task of localizing manipulation regions is similar to semantic segmentation task.

Given an image $X_i^m$, we pass it through the encoder to obtain $FM_i^m$ from one layer before the last convolutional layer. As we are interested in detecting manipulations in expression, we inject features from the facial expression recognition stream into the encoder-decoder manipulation detection stream. To do that, we also pass $X_i^m$ through FER and obtain features $FF_i^m$. We then concatenate both the feature maps as $F_i = FM_i^m \oplus FF_i^m$ and pass the concatenated feature maps through remaining layers of encoder to obtain latent space features. As shown in Fig. 2.2, we have two loss functions for classification ($\mathcal{L}_{cls}$) and segmentation ($\mathcal{L}_{seg}$). We use cross-antropy loss function for classification task defined as follows:

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_i \| y_i^m \log(a_i) + (1 - y_i^m) \log(1 - a_i) \|_1, \tag{2.6}$$

where $a_i$ is the output of binary classification which determines whether or not an image is manipulated.

Next, the decoder takes the latent space features as the input. Consider that $S_i \in \mathbb{R}^{H \times W}$ is the spatial manipulation segmentation output. We compute the segmentation loss function to measure the agreement between the segmentation mask and the ground-truth mask as follows:

$$\mathcal{L}_{seg} = \frac{1}{N} \sum_i \| M_i^m \log(S_i) + (1 - M_i^m) \log(1 - S_i) \|_1, \tag{2.7}$$

Note that each pixel in $S_i$ lies in between 0 and 1 depicting the probability of it being manipulated or not.

The total loss function we optimize to learn the encoder decoder architecture is as follows:

$$\mathcal{L}_{MANI} = \mathcal{L}_{cls} + \mathcal{L}_{seg} \tag{2.8}$$

### 2.3.5 Overall Algorithm

Here we discuss the overall training strategy for EMD algorithm. This is presented in Algorithm 1. Consider that the FER network is parameterized by $\phi$ and the the encoder-decoder for manipulation detection is parameterized by $\theta$. We learn them separately. First we sample images from the facial expression dataset $\mathcal{X}^F$, compute the loss $\mathcal{L}_{FER}$ and update $\phi$ using it. We then sample images from the manipulated images dataset $\mathcal{X}^M$, pass them through both pretrained FER stream and encoder-decoder stream, compute the loss $\mathcal{L}_{MANI}$ and then update $\theta$.

**Algorithm 1** Overall EMD Algorithm

**Inputs:** 1. Expression Recognition Dataset: $\mathcal{X}^F$
2. Expression Manipulation Dataset: $\mathcal{X}^M$
**Output:** Manipulation Detection Network: $\theta$
**Random Init.:**
1. Facial Expression Recognition Net: $\phi$
2. Manipulation Detection Net: $\theta$
**while** *notconverged* **do**
  Mini-batch $B^f = \{X_i^f, y_i^f\}_{i=1}^B \sim \mathcal{X}^F$
  Compute: $\mathcal{L}_{FER}(B^f; \phi)$
  Update: $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}_{FER}$
**while** *notconverged* **do**
  Mini-batch $B^m = \{X_i^m, M_i^m, y_i^m\}_{i=1}^B \sim \mathcal{X}^M$
  Compute: $\mathcal{L}_{MANI}(B^m; \theta, \phi)$
  Update: $\theta \leftarrow \theta - \eta' \nabla_\theta \mathcal{L}_{MANI}$

## 2.4 Experiments

In this section, we perform extensive experiments on three benchmark datasets from FaceForensics++ [103] to investigate the efficacy of the proposed method. We show results on two datasets (Face2Face and NeuralTextures) where the images correspond to facial expression manipulation and also two datasets (DeepFake and DFDC) where the images undergo an identity change.

### 2.4.1 Datasets

**FaceForensics++ Dataset.** For our experiments, we used the videos offered by FaceForensics++ [104] [1]. FaceForensics++ (FF++) contain 1,000 real videos and 1,000 Fake videos for each type of manipulation including Face2Face (F2F), DeepFake (DF) and NeuralTextures (NT). For each category of real/fake videos, the dataset was split into 720 videos for training, 140 for validation, and 140 for testing. We used videos with light compression (quantization = 23) and high compression (quantization = 40). Images were extracted from videos using the settings in [24]: 200 frames of each training video were used for training, and 10 frames of each validation and testing video were used for validation and testing respectively.

**DeepFake Detection Challenge Dataset (DFDC).** Deepfakes are a recent off-the-shelf manipulation technique that allows anyone to swap two identities in a single video. In addition to deepfakes, a variety of GAN-based face swapping methods have also been published. The problem of deepfake detection has received considerable attention, and this

[1] https://github.com/ondyari/FaceForensics

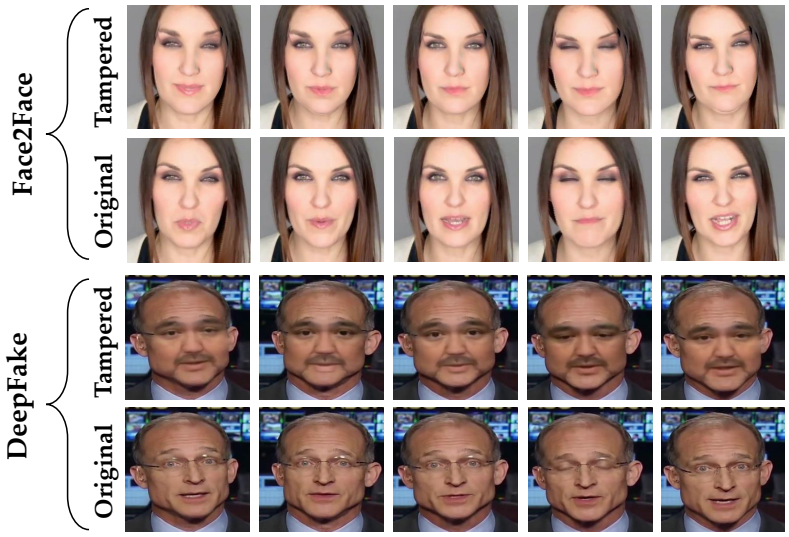



Figure 2.3: Two examples of pristine videos and their manipulated versions from F2F and DF datasets. As we can see, facial expression is manipulated in F2F videos while in DF datasets identities are swapped.

Table 2.1: Benchmark Datasets for face manipulation. Our focus in this chapter of thesis is on second row, i.e., detecting expression manipulations, while not degrading performance on the first row (identity manipulations).

| **Manipulation** | **Method** | **Dataset** |
|---|---|---|
| Identity Swap | FaceSwap, | UADFV, DF-TIMIT, |
| | DeepFakes | DFD, CelebDF, DFDC, Deeper Forensics 1.0, FF++ (FS and DF) |
| Expression Swap | Face2Face, | FF++ (F2F) |
| | NeuralTextures | FF++ (NT) |

research has been stimulated with many datasets. The DFDC dataset [31], different from the DF dataset as part of FF++ described above, is by far the largest publicly-available face swap video dataset with over 100,000 total clips sourced from 3,426 paid actors, produced with several deepfake, GAN-based, and non-learned methods.

We summarize the benchmark datasets for face manipulation in Table 2.1. As Table 2.1 shows, all the deepfake datasets are based on face identity swapping. The only

dataset that contains both identity and expression manipulation is FF++ [103]. In FF++, F2F and NT are the only manipulation techniques that change the facial expression, while two other techniques (FaceSwap and DeepFake) are based on identity change.

To demonstrate the difference between identity swap and expression swap, we show some examples from both categories. Fig. 2.3 shows 5 frames from 2 different face manipulation datasets (F2F and DF). As we can see, tampered videos from F2F are undergo expression manipulation. The shape of lips and eyebrows which contribute substantially to facial expressions is changed in most of the frames from F2F vidoes. To the contrary, the tampered videos from DF do not demonstrate any major expression change in comparison to original ones. Therefore, only two datasets (F2F and NT from FF++) satisfy the criteria and we evaluate the performance on those datasets. We also present the results on DF and DFDC datasets to demonstrate that our method performs at-par with the state-of-the-art methods in cases where the expression is not manipulated, but rather the identity is changed, thus ensuring generalizability of the approach.

**Facial Expression Datasets.** We use AffectNet [88], a new database of facial expressions. We trained our FER system on AffectNet training set and used F2F, NT, DFDC and DF datasets as manipulation detection datasets. AffectNet contains more than 1M facial images collected from the Internet. The dataset is divided into 11 facial expression categories - neutral, happiness, sadness, surprise, fear, disgust, anger, contempt, none, uncertain, and no-Face.

### 2.4.2 Implementation

In face forensics, faces play an important role and contain key features for manipulation detection. Therefore, instead of using the whole image, we extract the faces as a pre-processing step and only use the face regions to train the models. The FER system has shared and ensemble layers consisting of CNNs using 5x5 and 3x3 convolutional windows with the stride of 1. Following each convolutional layer is a batch normalization layer [51].

The encoder-decoder architecture consists of XceptionNet with separable convolution as encoder, spatial pyramid pooling and CNNs with 3x3 convolutional windows as decoder. We transfer XceptionNet to our task by replacing the final fully connected layer with two outputs. The other layers except last convolutional layer (the layer after feature concatenation) and fully connected layer are initialized with the ImageNet weights. To set up the inserted fully connected layer, we fix all weights up to the last convolutional layer and pre-train the network for 3 epochs. After this step, we train the network for 20 more epochs and choose the best performing model based on validation accuracy.

The framework is implemented on PyTorch. We trained the network using the ADAM optimizer [58] with a learning rate of 0.001, a batch size of 16, $\beta$ of 0.9 and 0.999, and $\epsilon$ equal to $10^{-8}$. The implementation consumes 11GB memory GPU and it takes 126hrs ( 5 days) for 20 epochs. For Encoder-Decoder (deeplabv3+ with pretrained Xception, fixed parameters) memory consumption and training time is the same with doubling of the batch size.

### 2.4.3 State-of-the-art Methods

**Manipulation detection.** Manipulation detection is a classification problem where the methods focus on identifying whether or not an image or video is manipulated. Some of the prominent works in this field that we compare against are as follows:

- **Steg. Features+SVM** [38] is based on steganalysis and employs handcrafted features. The features are co-occurrences on 4 pixels patterns along the horizontal and vertical direction on the high-pass images for a total feature length of 162. These features are then used to train a linear Support Vector Machine (SVM) classifier.

- **Cozzolino et al.** [24] combined the hand-crafted steganalysis features from [38] with a CNN-based network.

- **Bayar and Stamm** [12] proposed a CNN that uses a constrained convolutional layer followed by two convolutional, two max-pooling and three fully-connected layers. The constrained convolutional layer is specifically designed to suppress an image's content and learn manipulation detection features.

- **Rahmouni et al.** [97] proposed a CNN architecture with a custom pooling layer to optimize current best-performing algorithms' feature extraction scheme.

- **MesoNet** [6] is a CNN-based network using InceptionNet [56] in face forensics area. The network has two inception modules and two convolution layers with max-pooling, followed by two fully-connected layers. Instead of the cross-entropy loss, the authors use the mean squared error between true and predicted labels.

- **XceptionNet** [23] uses a deep neural network trained on ImageNet. This architecture is constructed by modifying the inception modules where the depthwise separable convolution is used. There are a total of 36 convolutional layers used in the base network. [104] transfered it to the manipulation detection task by replacing the final fully connected layer with two outputs. The other layers are initialized with the ImageNet weights.

- **LAE** [33], **FT-res** [25] both attempting to learn intrinsic representation instead of capturing artifacts in the training set.

- **DCNN** [96] proposes an approach leveraging the transferable features from a pre-trained Deep Convolutional Neural Networks (D-CNN) to detect manipulation in face images.

- **Two-stream** [143] uses a two-stream CNN to achieve higher performance in image forgery detection. They use standard CNN network architectures to train the model.

- **Capsule-Forensics** [91] presents a method that uses a capsule network to detect forged images and videos.

- **Face X-ray** [66] proposes a face forgery detection method based on the observation that most existing face manipulation methods share a common blending step and there exist intrinsic image discrepancies across the blending boundary.

**Manipulation Segmentation** In order to spatially localize the exact position of manipulation in an image, which in this case is part of a face, we use a decoder in our proposed architecture to segment out the tampered region. FaceForensics [103] and FaceForensics++

[104] datasets provide binary mask to show the manipulation region in the facial images. The most closely related work which also performs segmentation with classification is the following.

- **MultiTask** [90] outputs both the probability of an image being spoofed and segmentation maps of the manipulated regions in each frame of the input. For this method a CNN is designed that uses the multi-task learning approach to simultaneously detect manipulated images and videos and locate the manipulated regions for each query. Information gained by performing one task is shared with the other task.

### 2.4.4 Quantitative Comparisons

**Evaluation Metrics.** In terms of evaluation metrics, we use classification accuracy for the manipulation detection, which represents how many test images are correctly classified. For segmentation tasks, we use 1) pixel-wise classification accuracy which indicates whether a pixel in an image is manipulated or not, and 2) IoU (Intersection over Union). The IoU is calculated for both foreground and background, and the two IoUs are averaged to get mean IoU (mIoU).

**Expression Swap.** Table 2.2 shows the classification accuracy for the Face2Face (F2F) and NeuralTextures (NT) datasets (with expression swap) using two types of video quality (low quality (LQ) and high quality (HQ)). As may be observed, in terms of classification accuracy, EMD (our method) reaches the best performance on both datasets. In comparison to XceptionNet, our proposed method achieves $\sim 3\%$ and $\sim 2\%$ improvements in classification accuracy on F2F and NT datasets respectively. We also add FER to Mul-

Table 2.2: Classification performance in terms of accuracy for state-of-art architectures using two types of face manipulation including Expression Swap (F2F and NT datasets) and Identity Swap (DF and DFDC datasets).

| | **Method** | **Expression Swap** | | | | **Identity Swap** | | |
|---|---|---|---|---|---|---|---|---|
| | | F2F (HQ) | NT (HQ) | F2F (LQ) | NT (LQ) | DF (HQ) | DF (LQ) | DFDC |
| W/O FER | Steg.Features+SVM [38] | 74.68 | 76.94 | 60.58 | 60.69 | 77.12 | 65.58 | |
| | Cozzolino et al [24] | 85.32 | 80.60 | 62.08 | 62.42 | 81.78 | 68.26 | |
| | Bayar and Stamm [12] | 94.93 | 86.04 | 76.83 | 72.38 | 90.18 | 80.95 | |
| | Rahmouni et al [97] | 93.48 | 75.18 | 67.08 | 62.59 | 82.16 | 73.25 | |
| | MesoNet [6] | 95.84 | 85.95 | 83.56 | 75.74 | 95.26 | 89.52 | |
| | MultiTask [90] | 92.77 | 88.05 | 82.31 | 80.67 | 93.92 | 85.77 | 69.76 |
| | XceptionNet [104] | 98.23 | 94.50 | 91.56 | 82.11 | 98.85 | 94.88 | 88.98 |
| W/ FER | MultiTask+EnsFER | 95.22 | 89.15 | 85.89 | 81.46 | 94.10 | 86.31 | 70.02 |
| | EMD (ours) | 99.03 | 96.31 | 94.45 | 83.67 | 99.13 | 95.28 | 89.16 |

tiTask [90] architecture which leads to improvements of accuracy by $\sim 3\%$ and $\sim 1\%$ on F2F and NT datasets with high quality videos. Furthermore, we compare our method with more of the state-of-art methods on Face2Face dataset in terms of classification accuracy. Table 2.3 shows this comparison. As it is clear from Table 2.3, our method achieves higher classification accuracy. For localization task, Table 2.4 shows $\sim 3\%$ improvement of segmentation accuracy on low quality videos from F2F dataset and $\sim 2\%$ imporvement on NT dataset with the same video quality. We also achieved $\sim 5\%$ and $\sim 4\%$ improvement in mIoU on F2F and NT with low quality videos.

Table 2.3: Classification performance in terms of accuracy for state-of-art architectures on Face2Face datasets with two level of video quality.

| | **Method** | HQ | LQ |
|---|---|---|---|
| W/O FER | LAE [33] | 90.93 | - |
| | DCNN [90] | 93.50 | 82.13 |
| | FT-res [25] | 94.47 | - |
| | Two-stream [143] | 96.00 | 86.83 |
| | Capsule-Forensics [91] | 97.13 | 81.20 |
| | Face X-ray [66] | 97.73 | - |
| W/ FER | MultiTask+EnsFER | 95.22 | 85.89 |
| | EMD (ours) | 99.03 | 94.45 |

**Identetity Swap.** Table 2.2 also shows the classification accuracy for deepfake datasets where the identity is changed. We run experiments on two deepfake datasets including DFDC dataset [31] and DF [5]. Based on the results, our method achieves 89.16 % on DFDC while XceptionNet, as one the state-of-the-art methods, achieves 88.98 % in terms of classification accuracy. On DF dataset [5], we observe that with the addition of FER system, there is no fall in the performance. Thus, our method performs at-par with

Table 2.4: Segmentation performance in terms of accuracy and mIoU for all evaluated architectures on Face2Face and NeuralTextures datasets with two level of video quality.

| | **Method** | Acc(HQ) | | Acc(LQ) | | mIoU(HQ) | | mIoU(LQ) | |
|---|---|---|---|---|---|---|---|---|---|
| | | F2F | NT | F2F | NT | F2F | NT | F2F | NT |
| W/O FER | MultiTask [90] | 90.27 | 88.67 | 87.76 | 84.55 | 81.02 | 73.33 | 74.21 | 70.68 |
| | XceptionNet [104] | 96.13 | 91.34 | 92.45 | 89.39 | 89.71 | 79.25 | 81.47 | 75.55 |
| W/ FER | MultiTask+EnsFER | 93.22 | 90.56 | 89.31 | 86.56 | 83.25 | 77.46 | 78.33 | 74.23 |
| | EMD (ours) | 98.43 | 93.78 | 95.22 | 91.54 | 92.13 | 83.21 | 86.71 | 79.44 |

Table 2.5: Classification (cls) and segmentation (seg) accuracy for different FER architectures on Face2Face and NeuralTextures datasets with high quality videos.

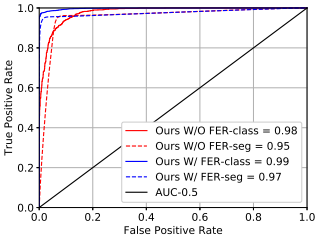
| **Method** | Face2Face | | NeutalTexture) | |
|---|---|---|---|---|
| | Cls | Seg | Cls | Seg |
| MultiTask+ SimFER | 94.93 | 91.84 | 88.62 | 89.21 |
| XceptionNet+ SimFER | 98.63 | 96.89 | 95.83 | 92.44 |
| MultiTask+EnsFER | 95.22 | 93.22 | 89.15 | 90.56 |
| EMD (ours) | 99.03 | 98.43 | 96.31 | 93.78 |

the state-of-the-art methods in cases where the expression is not manipulated, but rather the identity is changed. *This demonstrates the generalizability of our approach.*

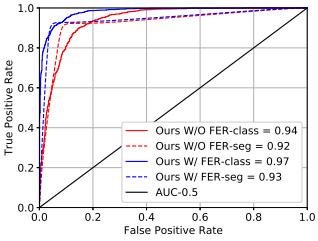
**Ablation study** To demonstrate the effectiveness of utilizing FER in manipulation detection and segmentation, we run different experiments with variation of FER architecture. As we can see from Table 2.5, using FER system with multiple branches and selecting the most informative feature maps by using ESR (the one we use in our architecture) [111] achieves higher accuracy in both detection and segmentation tasks. Using simple FER (SimFER) consisting of shallow convolutional layers (without Ensemble layers) leads to performance drop by $\sim 1\%$ and $\sim 2\%$ in classification and segmentation for both F2F and NT datasets with high quality videos.

### 2.4.5 Analysis of Results

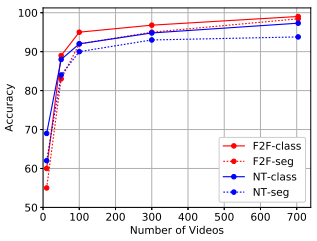
**Effect of FER on manipulation detection.** We use ROC curves to show the benefit of FER in manipulation detection. Figs. 2.4a and 2.4b demonstrate ROCs for both detection and segmentation tasks with and without FER system. AUC score for our network with FER stream (EMD) achieves 99% and 97% for detection and segmentation tasks on F2F respectively. Thus, our method leads to $\sim 1\%$ and $\sim 2\%$ improvement in




(a) ROC curves for classification and segmentation with and without facial expression recognition on F2F. The solid and dotted blue lines are the proposed EMD algorithm.




(b) ROC curves for classification and segmentation with and without facial expression recognition on NT. The solid and dotted blue lines are the proposed EMD algorithm.




(c) The detection and segmentation performance of our approach by varying the training corpus size on F2F and DF datasets.

Figure 2.4: Analysis of performance under different conditions.

detection and segmentation AUC score in comparison to its counterpart without the FER stream. Based on Fig. 2.4b, our method achieves $\sim 3\%$ and $\sim 1\%$ improvement in detection and segmentation AUC score when it is trained/tested on NT dataset.

**Size of training data.** As shown in Fig. 2.4c, we evaluate our proposed network on training sets with variable sizes. For both datasets (F2F and NT), we compute classification and segmentation accuracy varying the training size from 10 to $\sim 700$ videos. By adding more videos to the datasets our performance increases initially. Adding more than 300, our model's performance does not change much indicating our method can perform good enough when there is not much data to train with, i.e., $\sim 300$ videos.

**Why does FER help?** To show the effect of the features extracted from FER system, we visualize the last layer of CNN in our FER. For this purpose, we compute CAMs. A CAM for a particular category indicates the discriminative image regions used by the CNN to identify that category. Work by [142] has shown that the convolutional units

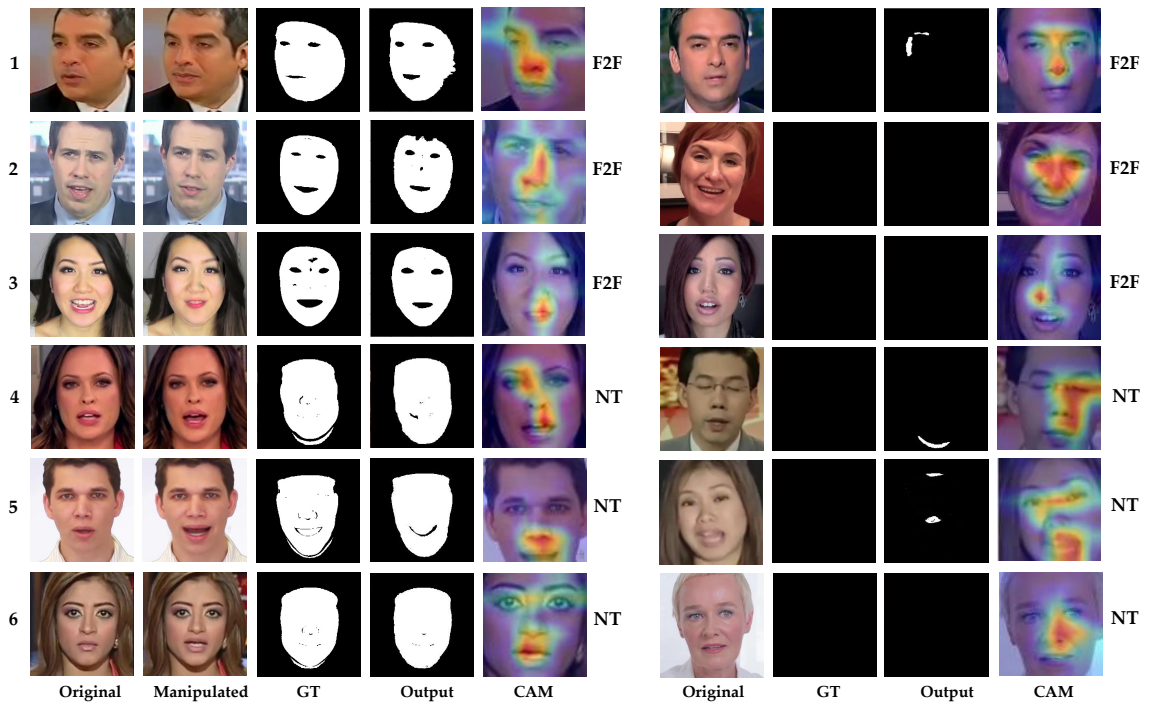



Figure 2.5: First and second columns of the left section show the original images and manipulated ones respectively. The black and white images in the third column are corresponding binary GT masks. Predicted masks (column 4) and generated class activation maps (column 5) for manipulated images from Face2Face (row 1,2,3) and NeuralTextures (row 4,5,6) dataset. In the right section of the figure, we show pristine images with their corresponding binary GT masks, predicted masks and CAMs.

of various layers of CNNs actually behave as object detectors despite no supervision on the location of the object provided.

In fact, the network can retain its remarkable localization ability until the final layer. This feature allows identification of the discriminative image regions which are important for manipulation detection. Specifically, for expression change detection, addition of a network which can localize regions in the face with information about the expressions helps manipulation detection methods to perform better.

As Fig. 2.5 represents, expression changes happen mostly around eyes, mouth and

eyebrows. In the last column of both right and left section of figure, we generate the CAMs for manipulated and pristine images in F2F and NT datasets. As it is clear, our network can classify expressions quite well although the main FER stream has been trained on a different dataset (AffectNet).

## 2.5 Additional Qualitative Results

**Datasets** In this section, we provide comprehensive information regarding datasets including expression and identity swap.

**Face2Face** Face2Face [118] is a facial reenactment system that transfers the expressions of a source video to a target video while maintaining the identity of the target person [104]. The modifications brought to the target image are in the form of change of movement of the head, lips, and facial expression.

**NeuralTextures** The NeuralTextures dataset [116] show facial reenactment as an example for their NeuralTextures-based rendering approach. It uses the original video data to learn a neural texture of the target person, including a rendering network. This is trained with a photometric reconstruction loss in combination with an adversarial loss. In the NeuralTextures, only the facial expressions corresponding to the mouth region is modified, when the eye region stays unchanged.

**DeepFake** DeepFake [5] is one of the popular face manipulation methods created using a Deep Neural Network, which swaps a facial image of a person with a different person's face, followed by editing [93].

**DeepFake Video Datasets.** Deepfakes are a recent off-the-shelf manipulation

Table 2.6: Benchmark Datasets for DeepFake Video Detection. Our approach is applicable to datasets that include facial expression manipulation. Only two datasets (Face2Face and NeuralTextures from Faceforensics++) satisfy that criteria.

| **Dataset** | **Released** | **# Videos** | | | **Real Video Source** | **Method** | **Fake Type** | |
|---|---|---|---|---|---|---|---|---|
| | | Real | Fake | Total | | | Id Swap | Exp Swap |
| UADFV [130] | Nov 2018 | 49 | 49 | 98 | YouTube | 1 | ✓ | |
| DF-TIMIT[61] | Dec 2018 | 0 | 620 | 620 | VidTIMIT | 2 | ✓ | |
| DFD [2] | Sep 2019 | 361 | 3070 | 3431 | YouTube | 5 | ✓ | |
| CelebDF [72] | Nov 2019 | 408 | 795 | 1203 | YouTube | 1 | ✓ | |
| DFDC [32] | Oct 2019 | 19154 | 99992 | 119146 | Actors | 8 | ✓ | |
| Deeper Forensics 1.0 [54] | Jan 2020 | 50000 | 10000 | 60000 | Actors | 1 | ✓ | |
| FaceForensics++ [103] | Jan 2019 | 1000 | 4000 | 5000 | YouTube | 4 | ✓ | ✓ |

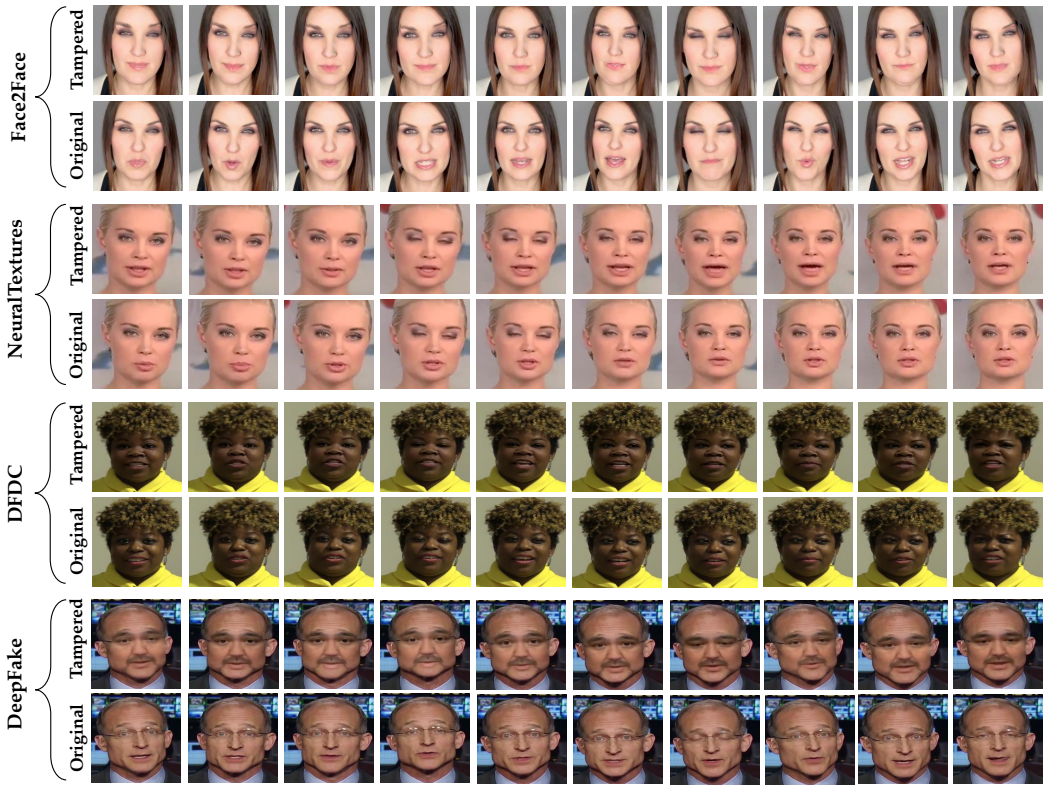



Figure 2.6: Four examples of pristine videos and their manipulated versions from F2F, NT, DFDC and DF datasets. As we can see, facial expression is manipulated in F2F and NT videos while in DFDC and DF datasets identities are swapped.

technique that allows anyone to swap two identities in a single video. In addition to Deepfakes, a variety of GAN-based face swapping methods have also been published. The problem of deepfake detection has increased considerable attention, and this research has been stimulated with many datasets. The DFDC dataset [31] is by far the largest currently and publicly-available face swap video dataset, with over 100,000 total clips sourced from 3,426 paid actors, produced with several Deepfake, GAN-based, and non-learned methods.

We summarize and analyze seven benchmark deepfake video detection datasets in Table 2.6. As Table 2.6 shows, all the deepfake datasets are based on face identity swapping.

The only dataset contains both identity and expression manipulation is Faceforensics++ [103]. In Faceforensics++, Face2Face and NeuralTextures are the only manipulation techniques change the facial expression while two other techniques (FaceSwap and DeepFake) are based on identity change. Thus, all the other datasets contain only identity manipulated faces. Only the part of Faceforensics++ allows us to analyze expression manipulation.

To demonstrate the difference between identity swap and expression swap, we show some examples from both categories. Fig. 2.6 shows 10 frames from 4 different face manipulation datasets (Face2Face, NeuralTextures, DFDC and DeepFake). As we can see, tampered videos from Face2Face and NeuralTextures are undergo expression manipulation. The shape of lips and eyebrows which contribute substantially to facial expressions is changed in most of the frames from Face2Face vidoes. To the contrary, the tampered videos from DFDC and DeepFake do not demonstrate any major expression change in comparison to original ones. Therefore, only two datasets (Face2Face and NeuralTextures from Faceforensics++) satisfy the criteria and we evaluate the performance on those datasets. We also present the results on DeepFake datasets to demonstrate that our method performs at-par with the state-of-the-art methods in cases where the expression is not manipulated, but rather the identity is changed, thus ensuring generalizability of the approach.

## 2.6 Conclusion

In this chapter of thesis, we propose a new approach (EMD) to exploit facial expression systems in image/video facial expression manipulation detection. Application of deep network layers rich in information about facial expressions improves the manipula-

tion detector by making it learn the useful features for facial expression transformation. Experiments on two challenging datasets demonstrate our method has better classification and segmentation performance in facial expression manipulation detection in comparison to state-of-art results. Also, our method is close to the state-of-the-art methods for other kinds of manipulation (identity swap) detection, thus ensuring generalizability.

# Chapter 3

# A Curriculum Learning Framework for Robust Image-Text Retrieval from Noisy Captions

With the rapid growth of multimedia data, image-text retrieval has become a pivotal task. In this thesis, we focus on learning image-text retrieval models utilizing image-text pairs, where the text is generated by an automatic captioning process without any requirement of tedious annotation effort. However, such generated descriptions are likely to be noisy (e.g.,wrong detection of objects, misinterpretation of the image concept) due to the use of automated captioning process. As a result, models trained with these automatically generated noisy captions will result in sub-optimal retrieval performance. Current strategies to mitigate the effect of noise are to consider additional pre-processing steps and filtering out the inconsistent words from the text. While such methods can detect inconsistencies in

the text, they are ineffective at identifying mismatches that require joint understanding of images and texts. Thus, robustness to noise in the learning process is a superior solution. To this end, this chapter of thesis proposes a robust image-text retrieval learning approach. Our proposed framework, **N**o**I**se-robust **C**ross-modal R**E**trieval (**NICE**), utilizes data-driven curriculum learning to up-weight the clean samples while discounting the noisy portion of the training data. The data-driven curriculum learning is supervised using external knowledge. Such sample weights, learned during training, compel the framework to learn from clean examples. In this thesis, we specifically focus on the scenario where the textual description is noisy due to automatic captioning process similar to Alt text generation. Extensive experiments are conducted on two image-text datasets to demonstrate the effectiveness of the proposed approach compared to multiple state-of-the-art methods for both image-to-text and text-to-image retrieval. Our proposed NICE can reach up to 7% absolute improvement in performance over the best baseline (OSCAR) on two COCO and Flickr30K datasets.

## 3.1 Introduction

With the rapid growth of multimedia data, image-text retrieval has become a pivotal task. Learning image-text retrieval model requires large sets of images with descriptions. However, the process of providing precise descriptions using human annotators is expensive and tedious. The availability of multimodal data on the web makes it attractive to crawl images with their corresponding alt texts (alternative text) [1] and use them to learn cross-modal retrieval models.

[1] alt text: Generally accompanies images on the webpages and intends to describe the nature or the content of the image in text format [109]

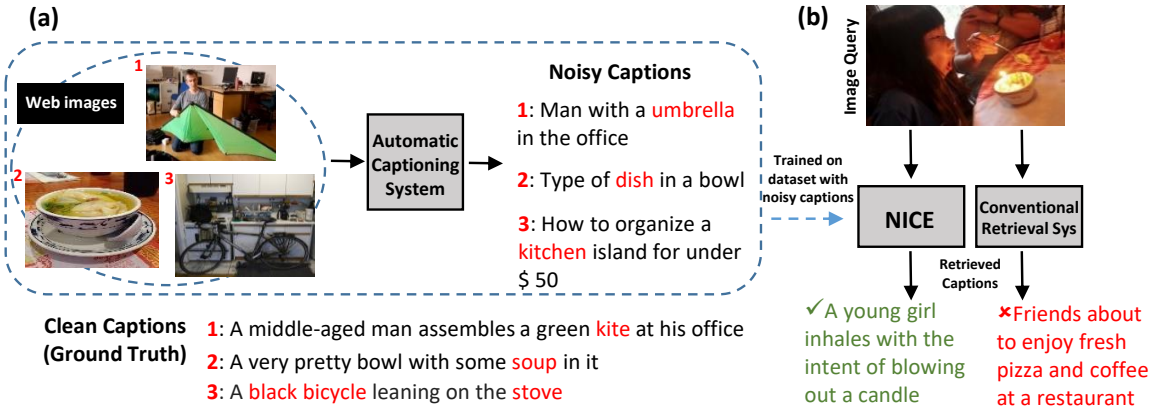



Figure 3.1: (a) Examples of generated noisy captions using image captioning model [82] pre-trained on Google Conceptual dataset [109]. Captions are generated by automatic image captioning system similar to alt text generation [107]. When automated captioning model is used to describe an image, the generated noisy captions are semantically meaningful but some of the details are not consistent with the image. (b) OSCAR [69], as one of the conventional retrieval systems, cannot retrieve the corresponding caption for an image from COCO [74] when it is trained on a noisy training set similar to Fig. 1a. Our proposed method (NICE) can successfully retrieve the corresponding caption.

Generally, alt text is automatically added when an image is uploaded and modern computer vision techniques are at the core of this automation process, e.g., Facebook's Automatic Alt Text (AAT) technology[2]. However, these texts are not in any way restricted or enforced to be accurate image descriptions. It can contain abstract information (e.g., search engine optimization terms, social media texts, hash-tags) and redundant information (e.g., file name, date updated, file identifier). It can also contain noisy/wrong descriptions of the images as the prediction/inference of automatic captioning process of different alt text generation technology can be inaccurate (check noisy captioning examples in Fig. 3.1 (a)).

Training an image-text retrieval model with such noisy image-alt text pairs would result in sub-optimal retrieval performance. One way to address the problem would be to

[2]https://www.facebook.com/help/ipad-app/216219865403298

manually identify and remove the noise from the text, which is itself a tedious and expensive task. Alternatively, we can apply text-based filtering approach similar to [52] to identify abstract and redundant information. These filtering methods are good at identifying inconsistencies that are solely in the text and can be used to get rid of abstract and redundant information present in alt text. However, they are unable to detect errors arising due to the failure of automatic captioning process such as in Fig. 3.1 (a), *which can only be detected by considering the image and text jointly.* In this thesis, **we specifically focus on the scenario where the textual description is noisy due to automatic captioning process**. Captions are generated by automatic image captioning system similar to alt text generation [107]. Our objective is to learn a robust image-text retrieval model utilizing image-text pairs where the text is generated by an automatic captioning system. To facilitate the experiments, we use a pre-trained captioning model to generate noisy captions for an unlabeled set of images. These image-text pairs are then used to train the robust retrieval system.

Noise in the captions leads to more difficulty in bridging the heterogeneity gap [137, 77] between image and text modality which is crucial for learning a cross-modal system. As is shown in Fig. 3.1 (b), OSCAR [69] (one of the state-of-the-art image-text retrieval systems) is not able to retrieve corresponding caption of an image when it is trained on images with noisy captions. This negative impact is empirically validated too (Fig. 4.3). Although numerous studies have been conducted to explore how to robustly learn from noisy data, such as [121, 47, 42, 43, 115, 131, 41, 85], *none of the works addresses the problem of learning robust image-text retrieval model in presence of noisy text annotation*

*generated by captioning model*. [47, 85] study cross-modal learning in presence of noisy labels. However, their setup has class label assigned to each image-text pair and only these class labels are assumed to be corrupted. Therefore, learning robust image-text retrieval model in presence of captioning system generated noisy text requires proper attention and thorough investigation.

To learn a retrieval model robust to noisy captions, it is imperative that the model is able to emphasize the clean part of the dataset and learn from it, while avoiding the noisy part. To this end, inspired by the success of Curriculum Learning (CL) to handle label noise [55], we propose a data-driven curriculum learning method using external knowledge. It learns to weigh the samples in a way that the model can focus on the samples whose captions have a high chance of being correct. For example, consider Fig. 3.1(a), which illustrates examples of noisy captions generated for COCO [74] images using a captioning model [82] pre-trained on Google Conceptual dataset [109]. The objects in the image and the words in the generated caption do not match (we see a kite in the image 1, while the caption has the word "umbrella" in place of "kite"). To identify the broken link between different modalities, we investigate the region and text embeddings in the joint embedding space. The misalignment between image regions and words can be detected in the joint embedding space if we have encoders to project the input in a way that relevant embeddings of different modalities remain close to each other while irrelevant ones are pushed far away. Motivated by this, we use a region encoder and a label encoder trained on an external source of data with correct annotation to obtain an approximation of how likely an image-text pair is correct. This approximation is used to guide the data-driven curriculum learning process

which can generate sample weights to learn from clean captions instead of corrupted ones. While some generated captions consist of multiple corrupted words, others are close to the original clean examples. Thus, learning from the clean part of the dataset becomes feasible even in presence of both the clean and noisy captions. We term our method as **N**o**I**se-robust **C**ross-modal R**E**trieval (**NICE**). From Fig. 3.1(b), we can see that our proposed method (NICE), trained in such a way, is able to retrieve the correct caption.

**Main contributions.** Followings are the main contributions of the work:

- We propose a novel problem of learning image-text retrieval model in presence of noisy captions, where the captions are generated by automatic image captioning system. Detecting such noise in the captions will requiring jointly understanding the correlations between the text and image features.
- We develop a novel framework for robust image-text retrieval systems. It utilizes data-driven curriculum learning to assign weights to the noisy image-text pairs depending on the caption reliability. Here, the data-driven curriculum is guided using external knowledge.
- Extensive experiments on two baseline datasets (COCO [74] and Flickr30k [132]) show that our proposed approach NICE brings significant improvement in performance compared to multiple state-of-the-art image-text retrieval methods in presence of noisy caption.

## 3.2 Related Work

**Image-text retrieval.** Multimodal learning involves the interaction between multiple modalities such as images and text. One line of multimodal learning focuses on image-text retrieval task. A common approach for image-text retrieval is to discriminatively embed the image and text to a shared visual-textual space[59, 35, 124, 140, 19, 34]. One of the early learning-based approaches[59] introduces an encoder-decoder pipeline that learns multimodal joint embedding space with images and text. Inspired by common loss functions used in structured prediction, [35] focuses on hard negatives for training. The authors in [124] use soft gate for fusion to adaptively control the information flow for message passing between the modalities.

With the emergence of transformers [120] in language understanding, recent works have explored the use of these architectures for vision and language tasks. Many such image-text [69, 87, 20, 63] or video-text[146] models rely on pre-trained object detectors to extract bottom-up features that are viewed as individual visual words. A few other works, such as PixelBERT [50] and VirTex [28] operate directly on dense feature maps instead of relying on object detectors. All aforementioned works in image-text retrieval learn from clean data while our proposed retrieval system is capable of learning from a mixture of clean and noisy captions.

**Learning with Noisy Data.** Learning from noisy data has been investigated widely in recent years. One approach to alleviate corrupted labels is based on noise correction strategy presented in [43, 115, 131, 41, 73, 121], aiming at removing samples with suspicious labels or correct their noisy labels to corresponding true ones. In [43], authors

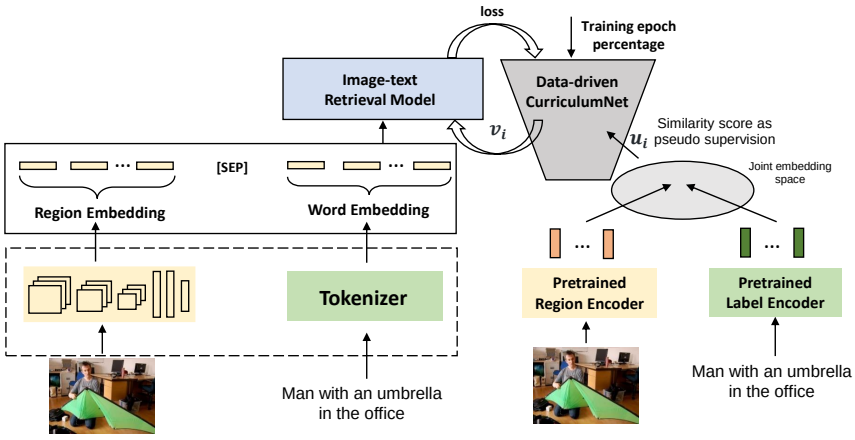



Figure 3.2: Our proposed framework (NICE) consists of three main components: i) the base image-text retrieval system, ii) the CurriculumNet, which estimates the importance weight of the image-text pairs in a data driven manner, and iii) a pre-trained region and label encoder that supervise the data-driven curriculum. The Image-text retrieval model and the CurriculumNet is trained in an iterative manner.

predict the pseudo-labels and train a deep network by using both the original labels and pseudo-labels. The proposed heuristic algorithms [115, 131] have the issue of removing too many instances or keeping mislabeled instances due to adjustment of a threshold for keeping or removing the samples. Authors in [73] propose a method to learn from noisy labels, by leveraging the knowledge learned from a small clean dataset and semantic knowledge graph to correct the noisy labels. The approach in [121] comprises a multi-task network that jointly learns to clean noisy annotations and to accurately classify images. The way these methods deal with noise require extra ground truth information indicating the presence of noisy labels. Providing ground-truth labels cannot always be guaranteed. Noise tolerant methods such as [55, 42, 133] focus on identifying true labeled samples from noisy training data. These methods to deal with noisy data are mostly applicable in other tasks, rather than multimodal learning.

**Image-Text Retrieval with Noisy Data.** Most prior works for noisy labels are specifically designed for unimodal scenarios. Although Works in [85, 47] investigated the problem of robust image-text retrieval, they consider each image-text pair is assigned with a label and the label is noisy. These works presented a two-step pre-processing method to obtain clean labels and feed them to cross-modal methods. To alleviate the influence of noisy samples, [47] reshaped the loss function to down-weight difficult samples and up-weight easy ones and thus focused training on clean samples. However, this method can not be extended to the case where the caption itself is noisy (where the words in a caption are swapped). Another strategy known to overcome noisy caption is filtering before training. Work in [52, 112] exploits some degree of text filtering for this purpose. However, these pre-proprosseing steps are expensive, time-consuming and ineffective at identifying mismatches between image and text features (see explanation earlier). In this chapter of thesis, we propose a method to learn directly from noisy captions without any pre-processing step.

## 3.3 Methodology

This section details our proposed approach for the task of image-text retrieval in presence of noisy captions. First, we define the problem statement (Section 3.3.1). Then we discuss the image-text retrieval model (Section 3.3.2) and the robust learning approach using data-driven curriculum and utilizing external knowledge (Section 3.3.3).

### 3.3.1 Problem Statement

Our objective is to learn a robust image-text retrieval system from a set of available images with noisy captions. Let $\mathcal{X}$ denote a set of $N$ images with noisy captions, i.e., $\mathcal{X} = \{(x_1, c_1), ..., (x_N, c_N)\}$, where $x_i$ is the $i^{th}$ image and $c_i$ is its corresponding noisy caption. The trained robust image-text retrieval system should be able to retrieve the relevant correct caption based on the given image and text similarly, retrieve the relevant image based on the given caption.

### 3.3.2 Image-Text Retrieval Model

The base structure of our image-text retrieval model consists of Transformer Encoder with stacked self-attention layers and a linear layer; similar to OSCAR [69]. the linear layer is used as the classifier to identify if a given image-text pair is aligned or not. For an input image-text pair, proper identification of the alignment requires fused vision-language representation to be accurate. Formally, assume we have a set $\mathcal{X}$ of $N$ images with noisy captions, i.e., $\mathcal{X} = \{(x_1, c_1), ..., (x_N, c_N)\}$, where $x_i$ is the $i^{th}$ image and $c_i$ is its corresponding noisy caption. For each image, an object detector is used to obtain $t$, where $t$ is a sequence of words/labels representing high precision detected object tags in the image. As a result, for each image-caption pair $(x, c)$, input to our image-text retrieval model is a triplet $(x, c, t)$.

**Input Embeddings.** For an image $x$, we extract features of $k$ regions $\{r'_1, r'_2, \ldots, r'_k\}$ using Faster R-CNN [99]. For each region, we concatenate 4 dimensional bounding box information predicted by Faster R-CNN to obtain position aware features. A linear layer is

used to obtain projected region features $\{r_1, r_2, \ldots, r_k\}$, which has the same vector dimension as the word embeddings. We use discrete tokens comprised of vocabulary words to represent captions. These tokens are the semantic units for processing the language. For each words in a caption, we use its representative token. Then the embedding of each word is the sum of its token-specific learned embedding [30] and encodings for position (token's index in the sequence). Similar to words in a caption, each tag is also represented by its corresponding token-specific learned embedding added with the positional encoding. For each image-caption instance $(x, c)$, we combine the embeddings of image, text, and tags with embedings of some special tokens (classification token: `[CLS]`, separator token: `[SEP]`) sequentially following the format: '`[CLS] caption [SEP] tags [SEP] image`'. This sequence is input to the image-text retrieval model for each image-caption pair.

**Image-Text Retrieval Model.** The image-text retrieval model consists of transformer encoder and a classification layer. The transformer encoder $\mathcal{E} : \mathcal{X} \to \mathcal{R}^d$ maps image-caption pair $(x, c) \in \mathcal{X}$ to a fused vision-language representation $\boldsymbol{h} \in \mathcal{R}^{d_h}$ (final embedding corresponding to the `[CLS]` token of the input image-caption pair). Here, $d_h$ is the feature dimension of the fused vision-language representation $\boldsymbol{h}$. This fused representation $\boldsymbol{h}$ is used for classification task. A classifier $\mathcal{F} : \mathcal{R}^{d_h} \to \mathcal{R}$ is used to identify if the fused vision-text representation $\boldsymbol{h}$ corresponds to correct image-text pair. Combining the encoder $\mathcal{E}$ and classifier $\mathcal{F}$, we define the trainable image-text retrieval model $\mathcal{T}_\theta : \mathcal{X} \to \mathcal{R}$ parameterized by $\theta$. Here, $\mathcal{T}_\theta(x, c)$ outputs the probability of $(x, c)$ being the correct image-caption pair. **Our objective is to learn a robust $\mathcal{T}_\theta$ for image-text retrieval task.**

### 3.3.3 Data-driven Curriculum for Robust Learning

To train the image-text retrieval model $\mathcal{T}_\theta$, the retrieval task is formulated as a binary classification problem. For each given image-text pair $(x, c)$ with noisy caption $c$, we formulate three negative / unaligned image-text pairs by i) sampling two random captions for the given image $(x, c_1^-)$, $(x, c_2^-)$, and ii) sampling a random image given the caption $(x^-, c)$. So for a given image-caption pair $(x, c)$ with noisy caption $c$, we have a collection of image-text pairs $\boldsymbol{s}$ and a collection of labels $\boldsymbol{y}$ indicating if each of the pairs are aligned according to the provided supervision. We use the image-text retrieval model $\mathcal{T}_\theta$ to obtain the output probabilities of the image-caption pairs being aligned. Hence,

$$\boldsymbol{s} = \begin{bmatrix} (x, c) \\ (x, c_1^-) \\ (x, c_2^-) \\ (x^-, c) \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_{(x,c)} \\ y_{(x,c_1^-)} \\ y_{(x,c_2^-)} \\ y_{(x^-,c)} \end{bmatrix}, \quad \mathcal{T}_\theta(\boldsymbol{s}) = \begin{bmatrix} \mathcal{T}_\theta(x, c) \\ \mathcal{T}_\theta(x, c_1^-) \\ \mathcal{T}_\theta(x, c_2^-) \\ \mathcal{T}_\theta(x^-, c) \end{bmatrix}$$

To learn an image-text retrieval model that can predict the image-text alignment correctly, cross-entropy loss, $l(\cdot)$, between predicted alignment scores and provided ground truth information is used. The optimization problem is to find the model parameter that minimizes cross-entropy loss computed over all the image-caption samples:

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} L^{\mathcal{T}}\big(\boldsymbol{y}_i, \mathcal{T}_\theta(\boldsymbol{s}_i)\big) + \gamma ||\boldsymbol{\theta}||_2^2, \tag{3.1}$$

where $L^{\mathcal{T}}\big(\boldsymbol{y}, \mathcal{T}_\theta(\boldsymbol{s})\big)$ is the cross-entropy loss computed for all four samples in $\boldsymbol{s}$, i.e.,

$$L^{\mathcal{T}}\big(\boldsymbol{y}, \mathcal{T}_\theta(\boldsymbol{s})\big) = \frac{1}{4}\sum_{j=1}^{4} l\big(\boldsymbol{y}[j], \mathcal{T}_\theta(\boldsymbol{s})[j]\big) \tag{3.2}$$

However, in our problem setting, since the captions are noisy, an image-text retrieval model trained on provided ground truth information optimizing Eqn. 3.1 will result in sub-optimal performance. Our goal is to learn a robust model $\mathcal{T}_\theta$ with the provided noisy caption. Towards that goal, during training, the idea is to emphasize more on the images with correct caption and diminish the impact of incorrect captions. In that regard, inspired by [55], we adopt a data-driven curriculum learning strategy and design CurriculumNet ($\mathcal{Z}_w$) to estimate the sample weights. Curriculum learning [55] defines a scheme under which training samples will be gradually learned and similarly we want to define a scheme where the clean captions will be emphasized. However, unlike [55], we don't have a small set of instances with clean labels from the same distribution of the training data to guide the data-driven curriculum generation process. Instead, we use knowledge learned using external data from different distribution to supervise the data-driven curriculum. First we discuss how we utilize the external knowledge to obtain supervision for the data-driven curriculum and how we design and train the CurriculumNet to estimate the sample weights.

**Learning External Knowledge.** We use object regions and object labels of Visual Genome dataset [62] as the external source of knowledge. Using the data, we train a region encoder $\mathcal{I}$ and a label encoder $\mathcal{H}$ to project the region representation and label representation in the joint embedding space. We project object region embedding $\boldsymbol{r}''$ and

**Algorithm 2** NoIse-robust Cross-modal REtrieval (**NICE**)

**Input:** Image ($x_i$)-text ($c_i$) pair set $\mathcal{X}$ where the text is noisy.
**Output:** Robust image-text retrieval model parameters $\boldsymbol{\theta}$.
**Initialization:** Initialize image-text retrieval model parameters $\boldsymbol{\theta}$ and CurriculumNet parameters $\boldsymbol{w}$
**while** not converged **do**
    Fetch Mini-batch of image-text pairs $B^f = \{x_i^f, c_i^f\}_{i=1}^b \sim \mathcal{X}$
    Compute image-text allignment loss $L^{\mathcal{T}}(B^f; \theta)$ using Eq. (2)
    Update retrieval model parametrs by $\theta \leftarrow \theta - \eta \nabla_\theta L^{\mathcal{T}}$
    **if** update curriculum **then**
        Compute input feature $\phi(x_i, c_i)$ of CurriculumNet
        Compute $L^{\mathcal{Z}}(B^f; w)$ of Eq. (5) for predicted weights
        Update CurriculumNet parameters by $w \leftarrow w - \eta \nabla_w L^{\mathcal{Z}}$
**Return** Retrieval model parameters $\boldsymbol{\theta}$

it's corresponding label embedding $\boldsymbol{t}''$ in the joint embedding space and minimize the cosine distance between them. For this purpose, we optimize mean multiplication of region embeddings and word embeddings as

$$L^S = -\frac{1}{n}\frac{1}{k}\sum_{i=1}^{n}\sum_{j=1}^{k}(\log(\sigma(\boldsymbol{r}''_{i,j}.\boldsymbol{t}''^{pos\intercal}_{i,j})) + \log(-\sigma(\boldsymbol{r}''_{i,j}.\boldsymbol{t}''^{neg\intercal}_{i,j}))), \tag{3.3}$$

where, $n$ and $k$ is the number of images and number of regions for each image $x_i$ respectively. $\boldsymbol{t}''^{pos}_{i,j}$ refers to embeddings of correct labels and $\boldsymbol{t}''^{neg}_{i,j}$ is for the sampled incorrect labels. $\sigma$ is Sigmoid function. The learned region encoder $\mathcal{I}$ and label encoder $\mathcal{H}$ is used for guiding the curriculum learning.

**CurriculumNet.** CurriculumNet $\mathcal{Z}_w$ is a neural network consisting of a LSTM layer and two fully connected layers. For a mini-batch of samples, it outputs their corresponding sample weights $v$ which is used for training image-text retrieval model $\mathcal{T}_\theta$. We use the current loss $L^{\mathcal{T}}\big(\boldsymbol{y}_i, \mathcal{T}_\theta(\boldsymbol{s}_i)\big)$ from $\mathcal{T}_\theta$, loss difference to the moving average, and embed-

ding of epoch percentage as input to the CurriculumNet. To supervise the training of the CurriculumNet, we use learned region encoder $\mathcal{I}$ and label encoder $\mathcal{H}$. For an image-caption pair $(x, c)$, we generate region embedding $\boldsymbol{r}''$ from the image and parsed noun embedding $\boldsymbol{t}''$ using $\mathcal{I}$ and $\mathcal{H}$ respectively and compute distance of embeddings by taking mean of inner products of embeddings:

$$u = \frac{1}{k}\frac{1}{m}\sum_{j=1}^{k}\sum_{p=1}^{m}(\boldsymbol{r}''_j.\boldsymbol{t}''_p), \tag{3.4}$$

where $k$ is number of object region in the image and $m$ is number of word tokens in the corresponding caption. $u$ is an indicator of how likely the image-caption pair is correct and used as a pseudo supervision to train the CurriculumNet. For an image-caption pair $(x, c)$ the output of the curriculumNet is $v = \mathcal{Z}_w(\phi(x, c))$, where $\phi(x, c)$ is the input feature (current loss $L^{\mathcal{T}}\big(\boldsymbol{y}_i, \mathcal{T}_\theta(\boldsymbol{s}_i)\big)$ from $\mathcal{T}_\theta$, loss difference to the moving average, and embedding of epoch percentage) to CurriculumNet for the corresponding image-caption pair. We use cross-entropy loss $L^{\mathcal{Z}}\big(u, \mathcal{Z}_w(\phi(x, c))\big)$ between predicted weight $v$ and estimated weight $u$ from the external knowledge to optimize the weights of CurriculumNet by

$$\arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} L^{\mathcal{Z}}\big(u_i, \mathcal{Z}_w(\phi(x_i, c_i))\big) + \beta||\boldsymbol{w}||_2^2 \tag{3.5}$$

**Final Objective.** The overall training strategy for NICE algorithm is presented in Algorithm 2. Instead of optimizing Eqn. 3.1, we use data driven curriculum to assign different weights to different samples.

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} v_i L^{\mathcal{T}}\big(\boldsymbol{y}_i, \mathcal{T}_\theta(\boldsymbol{s}_i)\big) + \gamma||\boldsymbol{\theta}||_2^2, \tag{3.6}$$

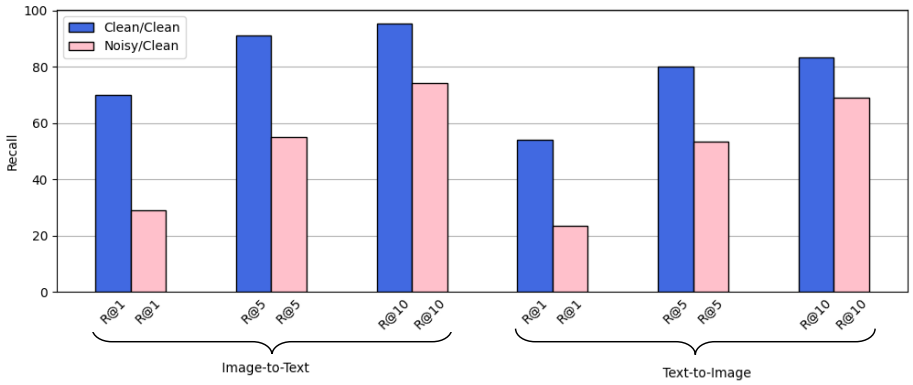



Figure 3.3: OSCAR [69] retrieval performance on COCO 5k [74] when train and test set both are clean (blue) vs when train set is noisy and test set is clean (pink). We see performance drop in Noisy training scenario.

where, $v_i = \mathcal{Z}_w(\phi(x_i, c_i))$ is the predicted weight by CurriculumNet. Since $\phi(x, c)$ gets updated by optimizing image-text retrieval model, we also optimize the CurriculumNet every other epoch for the current image-text retrieval model by,

$$\arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} L^{\mathcal{Z}}\big(u_i, \mathcal{Z}_w(\phi(x_i, c_i))\big) + \beta||\boldsymbol{w}||_2^2 \tag{3.7}$$

## 3.4 Experiments

In this section, we experimentally evaluate the performance of our proposed NICE approach. We first describe the experimental setups consisting dataset description, evaluation metrics, and the implementation details of the experiments. Finally, we report and analyze the results both quantitatively and qualitatively.

Table 3.1: This table shows the performance comparison for image-text retrieval task in COCO dataset when models are trained with noisy captions. Proposed NICE achieves significant improvement over other approaches.

| **Method** | **COCO 5k** | | | | | | **COCO 1k** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Image-to-Text** | | | **Text-to-Image** | | | **Image-to-Text** | | | **Text-to-Image** | | |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VSE++ [35] | 21.0 | 49.6 | 63.0 | 15.3 | 39.1 | 51.8 | 32.4 | 58.6 | 71.3 | 24.8 | 50.9 | 62.5 |
| ReasoningNet [87] | 25.4 | 52.7 | 70.6 | 22.9 | 51.9 | 66.6 | 36.9 | 65.1 | 76.8 | 32.8 | 56.9 | 72.7 |
| **NICE-T** | 29.0 | 58.9 | 76.9 | 26.1 | 55.5 | 71.9 | 41.3 | 69.8 | 81.1 | 37.1 | 60.4 | 76.1 |
| OSCAR [69] | 28.9 | 54.9 | 74.2 | 23.3 | 53.5 | 68.9 | 40.1 | 66.7 | 79.7 | 35.6 | 58.3 | 75.4 |
| **NICE** | **32.3** | **60.9** | **81.3** | **27.1** | **58.8** | **74.2** | **45.3** | **72.9** | **84.9** | **40.1** | **62.2** | **78.1** |

Table 3.2: This table reports the performance for image-text retrieval task in Flickr30K dataset when models are trained with noisy captions. Proposed NICE achieves significant improvement over other approaches.

| **Method** | **Image-to-Text** | | | **Text-to-Image** | | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VSE++ [35] | 25.2 | 54.9 | 66.4 | 15.9 | 44.4 | 53.7 |
| ReasoningNet [87] | 27.4 | 58.2 | 69.0 | 20.8 | 49.9 | 65.6 |
| **NICE-T** | 30.9 | 63.9 | 75.6 | 25.3 | 54.8 | 70.7 |
| OSCAR [69] | 31.7 | 61.3 | 79.5 | 25.1 | 59.9 | 74.8 |
| **NICE** | **36.8** | **68.3** | **86.8** | **30.3** | **66.5** | **80.3** |

### 3.4.1 Experimental Setups

**Datasets.** We employ two commonly-used image-text pair datasets, COCO [74] and Flickr30k [132], to evaluate our proposed method. COCO dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. It contains 164K images split into training (83K), validation (41K) and test (41K) sets. Every images contains roughly 5 text descriptions on average. Flickr30k is a standard benchmark for sentence-based image description. It contains 31,783 images collected from Flickr, together with 5 reference sentences provided by human annotators. We split the dataset into 1,000 test images, 1,000 validation images, and 29,783 training images.

To generate noisy captions, we exploit an automatic caption generator pre-trained on Conceptual Caption (CC) dataset [109]. The caption generator is the winner architecture in Google's Conceptual Challenge[3]. Captions are generated by automatic image captioning system similar to alt text generation [107]. Feeding the images of COCO [74] and Flickr30k [132] datasets to the caption generatorleads to generation of noisy captions. We trained our framework on COCO and Flickr30k train set with original images and the generated noisy

[3] `https://github.com/ruotianluo/GoogleConceptualCaptioning`

captions. We did the evaluation on clean COCO and Flickr30k test set (i.e., both image and corresponding caption are from the original datasets). We also trained two baselines adapted with curriculum learning setup (NICE-L and NICE-T, see **Robustness of NICE to Web-Crawled Noisy Datasets**) using a subset of CC dataset (100k). We evaluated the adapted frameworks on clean COCO. We extract features of image regions using Faster R-CNN [99] pre-trained on Visual Genome [62] dataset. We also use object regions and object labels of Visual Genome as the external source of knowledge to train region and label encoder. Visual Genome Dataset is an annotated image dataset and knowledge base that connects structured image concepts to language.

**Evaluation Metric.** We employ the Recall@K (R@K) metric for evaluating the retrieval performance for both image-to-text and text-to-image retrieval task. The R@K metric measures the percentage of queries for which our system is able to retrieve the correct item among the first K results. We consider K $\in \{1, 5, 10\}$. Please note that all the evaluations are done in testing sets with clean captions.

**Implementation Details.** We employ multi-layer transformer similar to pre-trained OSCAR [69] base model as the baseline for image-text retrieval task. The transformer is initialized with parameters of BERT base ($H = 768$). $H$ is the hidden size. The layers of transformer encoder are freezed up to the last layer. The bottom-up features for image regions are extracted using Fast-RCNN [99] pre-trained on Visual Genome dataset [62]. Our framework is implemented in PyTorch. All the experiments are trained for 40 epochs with a batch size of 8. We use ADAM [58] optimizer with initial learning rate of 0.00001. The decay factor in computing the loss moving average for curriculum learning is

set to 0.90. The experiments are conducted on an NIVIDIA 2080Ti GPU.

### 3.4.2 Result Analysis

The following experiments are conducted to evaluate the performance of our proposed NICE:

- We analyze the significance of the problem, i.e., existing image-text retrieval models incompetence to noisy text data.

- We compare the performance of proposed NICE approach in presence of noisy caption during training with other image-text retrieval approaches.

- We analyze the advantage of using data-driven curriculum learning over using fixed curriculum learning.

- We visualize sample embeddings generated using external knowledge and how we are utilizing it for pseudo supervision.

- We investigate our methodology performance using web-crawled noisy datasets to indicate the ability of our approach in robustness to such noisy captions.

- We show qualitative examples that verify the significance of the proposed NICE.

**Significance of the Problem.** In Fig. 3.3, we consider OSCAR model, fine-tuned for image-text retrieval task in COCO dataset for two settings: i) the training images are associated with clean captions (blue) and ii) the training images are associated with noisy/wrong captions (pink). Model trained with noisy captions has $17\% - 42\%$ performance drop compared to model trained with clean captions. We also observe the performance drop

Table 3.3: Performance comparison for fixed vs data-driven curriculum on COCO 5k dataset for different image-text retrieval approaches. Data-driven curriculum provides highest robustness compared to using fixed curriculum and also compared to vanilla models.

| **Method** | $\mathfrak{Z}_w$ | **Image-to-Text** | | | **Text-to-Image** | | |
|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VSE++ | - | 21.0 | 49.6 | 63.0 | 15.3 | 39.1 | 51.8 |
| NICE-L | ✗ | 23.1 | 52.9 | 66.9 | 16.9 | 45.4 | 54.0 |
| **NICE-L** | ✓ | **25.1** | **53.9** | **68.7** | **18.9** | **47.4** | **56.4** |
| ReasoningNet | - | 25.4 | 52.7 | 70.6 | 22.9 | 51.9 | 66.6 |
| NICE-T | ✗ | 27.9 | 55.9 | 74.2 | 24.9 | 52.8 | 69.1 |
| **NICE-T** | ✓ | **29.0** | **58.9** | **76.9** | **26.1** | **55.5** | **71.9** |
| OSCAR | - | 28.9 | 54.9 | 74.2 | 23.3 | 53.5 | 68.9 |
| NICE | ✗ | 30.6 | 58.1 | 78.2 | 25.7 | 55.5 | 71.9 |
| **NICE** | ✓ | **32.3** | **60.9** | **81.3** | **27.1** | **58.8** | **74.2** |

for different approaches (VSE++, ReasoningNet) and for different datasets. It justifies that image-text retrieval model trained with caption generated from automated captioning system are not robust to caption noise.

**Robustness of NICE.** We analyse the robustness of our proposed NICE approach learned using noisy captions for image-text retrieval task. Ours is the first work that proposed noise robust image-text retrieval system where labels for different modality data are not available. Hence, there is no other noise robust approach we can directly compare with. Instead, we consider three different types of state-of-the-art image-text retrieval models: i) **VSE++** [35] (Linear layer based model), ii) **ReasoningNet** [87] (Transformer [120] based model), and iii) **OSCAR** [69] (BERT [29] based pre-trained model) and train these models with noisy captions for comparison. To show the efficacy of our proposed approach, we adapt both Transformer based image-text retrieval model **(NICE-T)** and BERT based pre-trained image-text retrieval model **(NICE)** with our proposed data-driven cur-

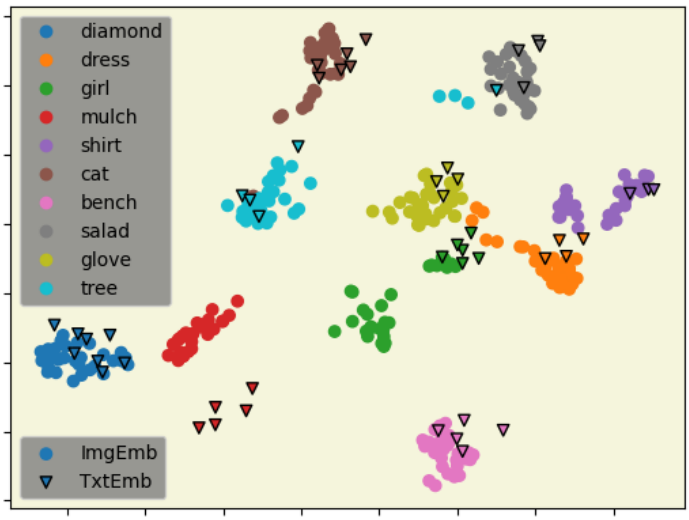



Figure 3.4: t-SNE visualization of region and label embeddings by $\mathcal{I}$ and $\mathcal{H}$. We observe small distances between text and image embedding of the same object; most of them are perfectly aligned.

riculum learning using external knowledge. Table 3.1 and Table 3.2 show the retrieval performance for COCO and Flickr30K dataset respectively. For both datasets, we observe that conventional image-text retrieval models perform poorly when trained with noisy captions. NICE-T provides $2\% - 6\%$ improvement over its base approach ReasoningNet for both datasets over all the reported metrics. Our proposed NICE results in $3\% - 7\%$ absolute improvement over the best baseline (OSCAR). It validates that NICE is robust to noise present in caption and the data driven curriculum learning using external knowledge is adaptable to different types of image-text retrieval system.

**Fixed Curriculum vs Data-driven Curriculum.** To make the image-text retrieval system robust to the noise in caption, our strategy is to emphasize more on the clean labels and diminish the impact of incorrect caption. We use external knowledge (pre-trained region encoder $\mathcal{I}$ and label encoder $\mathcal{H}$) to estimate how likely the caption is correct

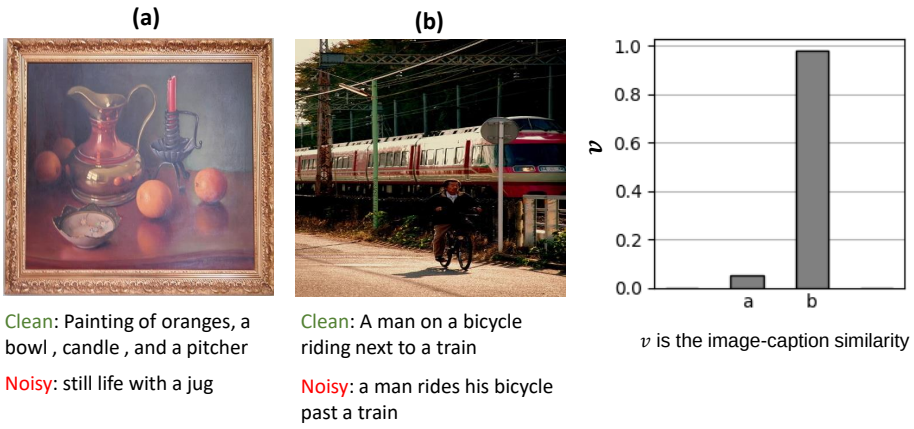



Figure 3.5: (a) Generated caption doesn't match with the clean caption which is successfully identified using external knowledge ($v$ is low). (b) Generated caption is successfully identified as correct ($v$ is high).

for an image-caption pair. Then we use this estimate to guide the CurriculumNet to generate sample weights during training. However, the estimate from the external knowledge can also be used directly to set up a fixed curriculum. In Table 3.3, we consider NICE with both fixed curriculum (curriculumNet $\mathcal{Z}_w$ not used) vs data-driven curriculum (curriculumNet $\mathcal{Z}_w$ is used). We observe that even when fixed curriculum is used, there is $2\% - 4\%$ improvement of performance compared to approaches that don't have any noise robust component. However, when data-driven curriculum is used, we obtain $3\% - 7\%$ improvement over the approaches that don't have any noise robust component in COCO dataset. Here, NICE-L, NICE-T, and NICE refer to linear layer based, transformer based, and BERT based image-text retrieval model adapted with curriculum learning setup respectively. Our intuition is that when CurriculumNet ($\mathcal{Z}_w$) generates sample weights based on the output of image-text retrieval model ($\mathcal{T}_\theta$), iterative training of these model provides robustness to each other. As a result, it performs better compared to using fixed curriculum and justifies the use of data-driven

Table 3.4: This table reports the performance for image-text retrieval task in COCO 5k dataset when models are trained with CC dataset. Proposed NICE-L and NICE-T achieves improvement over conventional approaches.

| **Method** | **Image-to-Text** | | | **Text-to-Image** | | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VSE++ [35] | 13 | 35.7 | 52.1 | 10.3 | 30.6 | 40.1 |
| **NICE-L** | 14.7 | 37.5 | 56.8 | 11.1 | 32 | 43.4 |
| ReasoningNet [87] | 16 | 38.1 | 57.1 | 12.3 | 33 | 44.3 |
| **NICE-T** | 17.9 | 40.1 | 62.8 | 13.1 | 35.5 | 47.9 |

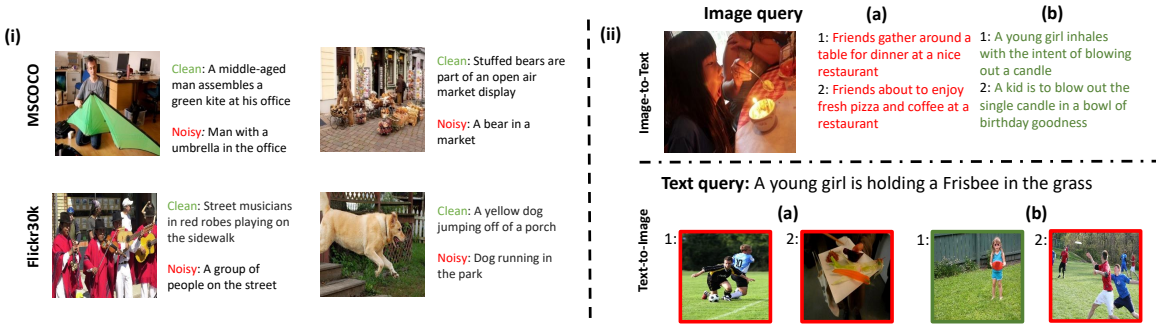



Figure 3.6: Qualitative retrieval results. (i) Examples of image-text pairs from COCO [74] and Flickr30k [132]. "Clean" refers to caption from the original dataset, and "Noisy" descriptions are generated noisy captions. (ii) The top-2 retrieved results are shown. Green denotes the ground-truth images or captions. Our model can capture the alignments between images and captions even though the training set consists of noisy captions.

curriculum.

**Analysis of Learned External Knowledge.** We use region encoder $\mathcal{I}$ and label encoder $\mathcal{H}$ learned from external source of clean data from different distribution, which guides the training of CurriculumNet $\mathcal{Z}_w$. Hence, learning a proper curriculum is dependent on the learned $\mathcal{I}$ and $\mathcal{H}$ from external source. To check the performance of $\mathcal{I}$ and $\mathcal{H}$, we visualize the learned semantic feature space of image-text pairs of Visual Genome [62] (clean dataset) on a 2D map using t-SNE [119] in Fig. 3.4. The embedding of the image and the word corresponding to the same object share the same color. As we can see,

the distance of the same object between two modalities is substantially low; most of them are perfectly aligned, as demonstrated by the overlapping regions. This shows learned $\mathcal{I}$ and $\mathcal{H}$ correctly projects object region and its label in the joint embedding space.

**Pseudo Supervision Using External Knowledge.** NICE uses learned knowledge from external source to compute how likely a image-caption pair is correct and use it as a pseudo supervision to train the CurriculumNet. For example, in Fig. 3.5, two images and their corresponding original and generated caption using automatic caption generator are given. As it is clear from Fig. 3.5 (a), the automatically generated caption is not a good description of the picture. As a result, the similarity score for (a) is low, which is an indicator that the information in the caption is not helpful to learn from. However, the caption for the image (b) in Fig. 3.5 is less noisy. Most of the words are the same as the original caption, while just the sentence structure is altered. So the similarity score for (b) is high and learning from this caption is beneficial. We exploit these similarity scores as pseudo supervision to train CurriculumNet.

**Robustness of NICE to Web-Crawled Noisy Datasets.** Even though the focus of the work is to learn a noise robust retrieval model specifically for the noisy texts generated by captioning models, we observe that our approach is also robust to the noise obtained from web crawled image-text data, similar to noisy captions in Conceptual Caption dataset [109]. To demonstrate robustness of our approach to such noisy dataset, we conducted experiments in which VSE++ [35] and ReasoningNet [87] are trained on CC. Table 3.4 demonstrates retrieval performance for subset of CC dataset (100K). We did not conduct the similar experiment for NICE because it is already pre-trained on CC dataset.

We evaluated VSE++ and ReasoningNet on test collections from MS-COCO which is more similar to CC. As seen in Table 3.3, NICE-L and NICE-T, which are linear layer based and transformer based image-text retrieval model adapted with curriculum learning respectively, perform quite well in comparison to VSE++ and ReasoningNet. The presented results indicate the fact that our methodology is also compatible with other types of noisy captions in image-text pair datasets.

**Qualitative Results.** We qualitatively demonstrate some image-text pairs from COCO [74] and Flickr30k [132] along with generated noisy captions. Fig. 3.6 (i) elaborates the fact that some generated captions consist of multiple corrupted words, while some other generated captions are close to clean ones. Thus, learning from the clean part of the dataset is feasible with the existence of correct captions and noisy ones. In Fig. 3.6 (i), the captions in the first line of the description are the original texts, and the second rows include the generated noisy captions.

Fig. 3.6 (ii) shows top-2 ranked image-to-text and text-to-image retrieval results of two samples from COCO [74]. In the image-to-text retrieval case (top sample), captions in column (a) are retrieved by OSCAR [69]. Red captions show the wrong retrieval. As it is clear, OSCAR, trained with noisy caption, is not capable of retrieving any captions corresponding to the image. In column (b), our proposed framework retrieves the correct captions marked in green. In comparison to OSCAR [69], our method has promising performance as it retrieves two correct captions. The below sample in Fig. 3.6 (ii) demonstrates the text-to-image case. The row (a) of images for the caption query is from OSCAR [69], and the row (b) is for our proposed method. While OSCAR fails to retrieve the images

corresponding to the query, our proposed method is able to retrieve results which connect the query and corresponding image.

## 3.5 Additional Qualitative Results

In this section, 1) we provide the complete pseudo-code of the our proposed method, Noise Robust Cross-modal Retrieval (NRCR). 2) we present the detailed architecture of our framework including the image-text retrieval system, data-driven curriculum net and region-label encoder using external knowledge. 3) we show more qualitative results for image-to-text and text-to-image retrieval. 4) we provide our generated noisy captions using captioning model [82].

**Algorithm** The detailed pseudo-code of the proposed paradigm is described in 3. It can be summarized into two-stage training schemes: region and label encoder training using external knowledge for stage one, image-text retrieval system training with robust learning approach using data-driven curriculum for stage two.

**Detailed Architecture** In page 2 we provide complete and detailed architecture of our proposed framework (NICE). Our framework consists of image-text retrieval model which is a multi-layer transformer. The output feature of [CLS] representation is used to classify if image and text are related or not. Data-driven curriculumNet is a neural network consisting of a LSTM layer , embedding layer and two fully connected layers. The loss, loss difference to the moving average, and epoch percentage are input to the CurriculumNet. The concatenated outputs from the LSTM and the embedding layers are fed into two fully-

**Algorithm 3** **N**o**I**se-robust **C**ross-modal R**E**trieval (**NICE**),

Step 1: training region and label encoder using external knowledge $\mathcal{X}^{\mathcal{S}}$
**Input:** $\mathcal{X}^{\mathcal{S}} = \{r_i, j'', t_i, j''\}_{i=1,j=1}^{n,k}$ for n number of images and k number of regions.
**Output:** Region and label encoder ($\mathcal{IH}_\gamma$) parameters $\boldsymbol{\gamma}$.
Initialize region and label encoder parameters $\boldsymbol{\gamma}$
**while** Not converged **do**
  Fetch Mini-batch $B^S = \{x_i^S, c_i^S\}_{i=1}^B \sim \mathcal{X}^S$
  Compute: $L^S(B^S; \gamma)$ by Eq. (3)
  Update: $\gamma \leftarrow \gamma - \eta' \nabla_\gamma L^{\mathcal{S}}$

Step 2: training image-text retrieval system using image-text pairs with noisy caption $\mathcal{X}$
**Input:** Image-text pair set $\mathcal{X}$
**Output:** Image-text retrieval model parameters $\boldsymbol{\theta}$.
Initialize Image-Text retrieval model parameters $\boldsymbol{\theta}$,
Initialize CurriculumNet parameters $\boldsymbol{w}$
**while** Not converged **do**
  Fetch Mini-batch $B^f = \{x_i^f, c_i^f\}_{i=1}^B \sim \mathcal{X}$
  Compute: $L^{\mathcal{T}}(B^f; \theta)$ by Eq. (6)
  Update: $\theta \leftarrow \theta - \eta \nabla_\theta L^{\mathcal{T}}$
  **if** update curriculum **then**
    Compute $u_i = \mathcal{IH}_\gamma(x_i, c_i)$ where $\mathcal{IH}_\gamma$ is
    pretrained region and label encoder
    Compute $\phi'(u_i, \theta)$
    Compute: $L^{\mathcal{Z}}(\phi'; w)$ by Eq. (7)
    Update: $w \leftarrow w - \eta \nabla_w L^{\mathcal{Z}}$
**Return** $\boldsymbol{\theta}$

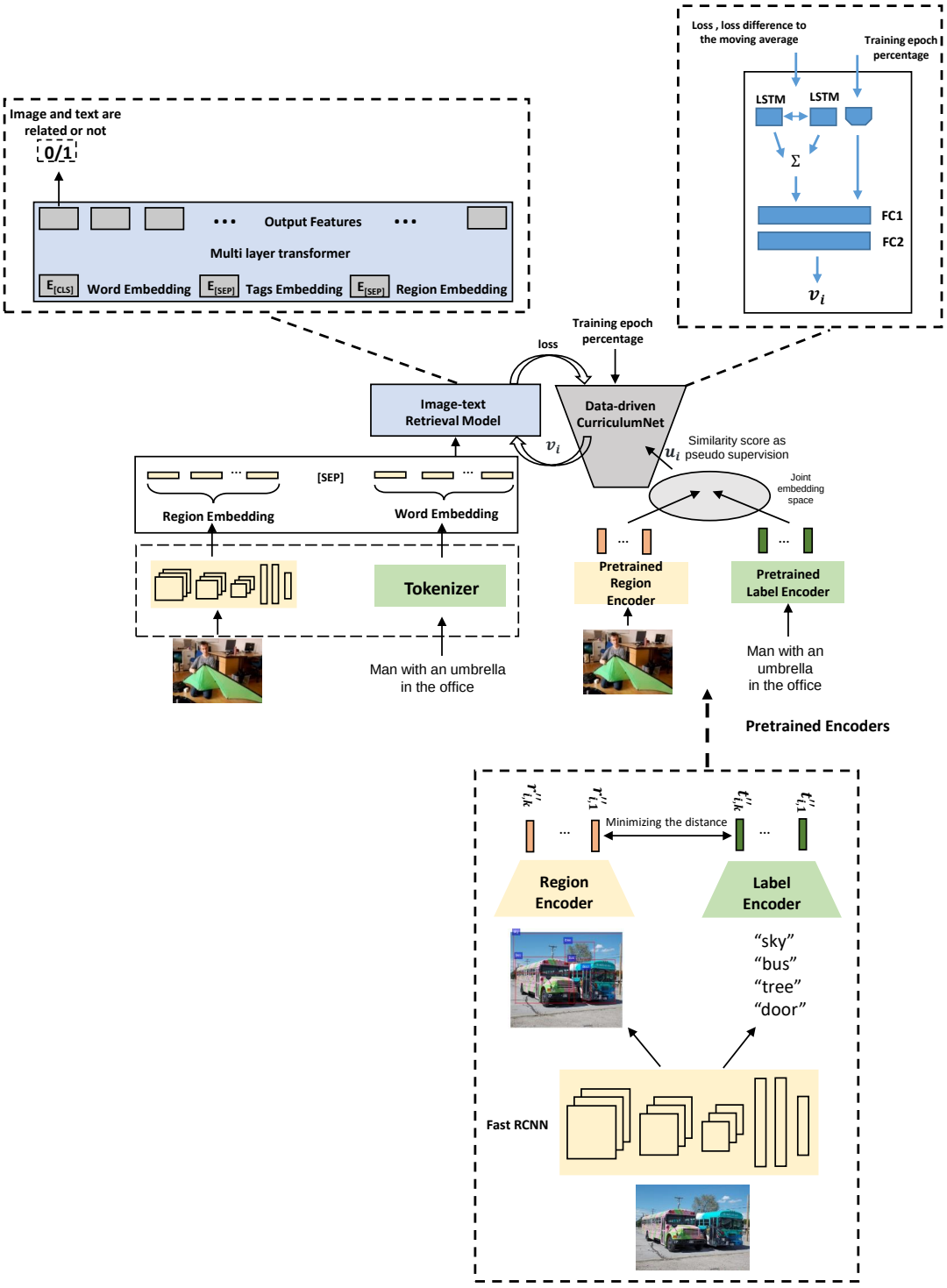



Figure 3.7: Detailed architecture of our proposed framework.

connected layers. We extract features of image region using Fast RCNN [99]. The extracted features along with labels are fed to region and label encoder to minimize the gap between the modalities.

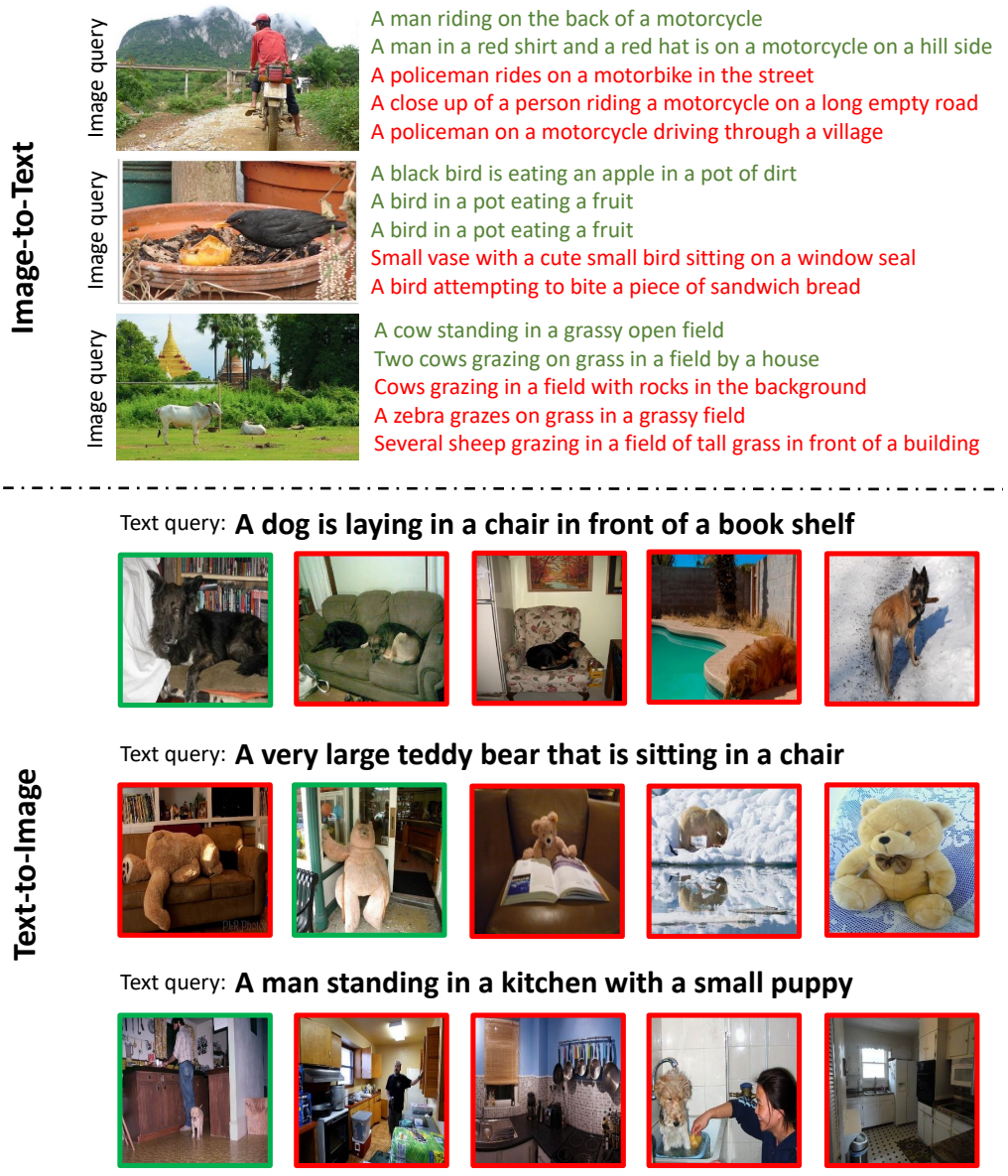



Figure 3.8: The top-5 retrieved results are shown. Green denotes the ground-truth images or captions

**Qualitative Results** We show top-5 ranked image-to-text and text-to-image retrieval results of six query samples from COCO [74]. Our image-text retrieval system is able to retrieve three correct captions for second image query in image-to-text retrieval scenario. As Fig. 3.8 shows in text-to-image scenario, the corresponding image shows up in first and second rank. Our proposed approach is capable of image-text matching although the trainset includes noisy captions. Green denotes the ground-truth images or captions.

## 3.6 Conclusions

In this paper, we propose a new image-text retrieval framework robust to noisy captions during the training process. This is an useful framework when we are learning from automatically generated captions, as is inevitable given the large volumes of data being collected. We exploit curriculum learning to up-weight the clean samples and prevent noise interference in the training process. To provide supervision for data-driven curriculum learning, we use an already annotated large dataset as an external knowledge source. Thus, our proposed framework is robust to noisy captions even though we have no access to a label indicating the sample is correct or not. Experiments on two challenging datasets demonstrate our method has better performance in comparison to state-of-the-art baselines. The paper addresses an important practical problem, how to learn cross-modal retrieval methods from imperfect captions, thus allowing scalability to large-scale automatically captioned image datasets. There is significant scope for improvement. Overall, cross-modal retrieval performance has a lot of room for improvement, and doing so from noisy captions has even more such scope.

# Chapter 4

# Learning Noise-Robust Image-Text Retrieval with Alignment Network

With the rapid growth of multimedia data, image-text retrieval has become a pivotal task. In this work, we focus on learning image-text retrieval models utilizing noisy image-text pairs. Contrary to prior works where the source of noise is the random shuffling of image-text pairs, we consider a realistic scenario where image captions were generated using an automatic captioning process without any tedious annotation effort. Since such generated descriptions are likely to be noisy (e.g., wrong detection of objects, misinterpretation of the image concept), models trained with these automatically generated noisy captions will result in sub-optimal retrieval performance. To this end, we propose a robust image-text retrieval learning approach. Our proposed method, Image-Text Robust Retrieval Alignment Network (ITRANE), utilizes an alignment model to bridge the gap between corresponding embeddings in different modalities. It enforces fine-grained alignment between vision and

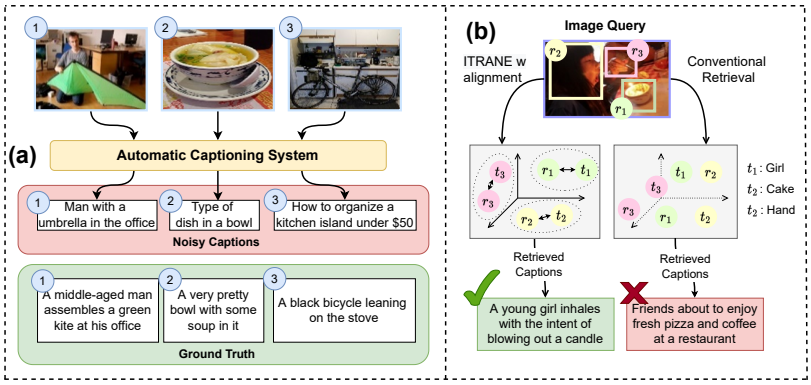



Figure 4.1: (a) Examples of generated noisy captions using captioning model [82]. When an automated captioning model is used to describe an image, the generated noisy captions are semantically meaningful but some of the details are not consistent with the image. (b) While a conventional retrieval system cannot retrieve the corresponding caption for an image when it is trained on a noisy training set similar to Fig. 1a, our proposed method, ITRANE, can successfully retrieve the corresponding caption since the alignment model enforces fine-grained alignment between corresponding embeddings in different modalities. $r$ and $t$ refer to the image region and the tags/labels detected by Faster R-CNN [99], respectively.

language modalities as a regularization constraint to increase the robustness of an image-text retrieval model. Extensive experiments are conducted on two image-text datasets to demonstrate the effectiveness and generalizability of ITRANE. Our results show that ITRANE can reach up to 14% average improvement in performance compared to a recent noise robust image-text retrieval approach.

## 4.1 Introduction

Cross-modal retrieval is the task of retrieving the relevant samples from one modality as per the given user query expressed in another modality. There has been a large body of work [50, 28, 69, 35, 140, 19, 34] that addresses the task of cross-modal image-text retrieval. Although these methods have shown promising performance, their success depends

on an implicit assumption of correctly annotated training data. However, the process of providing precise descriptions using human annotators is expensive and tedious. As a result, there can be scenarios where a system needs to learn image-text retrieval models using noisy image-text pairs.

Training an image-text retrieval model with noisy image-text pairs would result in sub-optimal retrieval performance. Few recent works [49, 94] have addressed the task of learning an image-text retrieval model utilizing noisy image-text pairs. They consider noisy correspondence, where the correct pairs were shuffled to create the noisy scenario. Such annotation does not reflect the practical scenario of web-crawled noisy image-text pairs or cheap noisy annotation from annotators. Instead, analogous to how Alt-texts [1] are automatically generated (e.g., Facebook's Automatic Alt Text (AAT) technology[2]), we use an automatic captioning model to generate the noisy description of images. This is a far more practically relevant model for noisy captions than shuffling them.

In this work, we specifically focus on the problem of learning a robust image-text retrieval system using noisy image-text pairs generated by a captioning model. One way to address the problem would be to apply a text-based filtering approach similar to [52]. These filtering methods are good at identifying inconsistencies that are solely in the text. However, they are unable to detect errors arising due to the failure of the automatic captioning process such as in Fig. 4.1 (a), *which can only be detected by considering the image and text jointly.* Few works [52, 112] scale up the training set to mitigate the effect of noise. However, scaling up the dataset may not be a feasible option due to resource

[1]Alt-text: Generally accompanies images on the webpages and intends to describe the nature or the content of the image in text format [109]

[2]https://www.facebook.com/help/ipad-app/216219865403298

constraints and in scenarios dealing with private data (e.g., healthcare) or hard-to-acquire data (e.g., environmental monitoring). Thus, we need robust approaches to learn from realistic noisy image-text pairs.

Prior works [84, 55, 42] have shown that training in presence of noisy annotation results in an over-fitted model to accommodate the noise. One way to tackle the issue is to use regularization techniques. However, common regularization techniques are not enough for generalization and cannot properly shield the effect of noisy data. Therefore, our proposed approach ITRANE considers an additional optimization problem as the regularization constraint to mitigate the effect of noise. Motivated by fine-grained image-text retrieval methods which decompose image-text matching into global-to-local level matching, we enforce fine-grained alignment between region embedding and text embedding as the regularization constraint. We leverage transformer-based image-text retrieval architecture, along with a readily available vision-language pretrained model (CLIP) and a pretrained object detector, to enforce fine-grained alignment. For example, in image ① of Fig. 4.1(a), we see a kite in the image, while the caption has the word "umbrella" in place of "kite". A pretrained object detector is likely to predict the tags/labels of the salient "kite" region correctly. We enforce the image region embedding of "kite" generated by an image-text retrieval model to be aligned with the text embedding of "kite" obtained from pretrained CLIP. Therefore, image region embeddings are learned close to correct corresponding tags/labels and pushed far away from irrelevant ones. This constraint guides the model to be robust to the presence of noise. From Fig. 4.1(b), we can see that ITRANE, trained in such a way, is able to retrieve the correct caption.

**Main contributions.** Followings are the main contributions:

- We address the problem of learning image-text retrieval models in presence of noisy captions, where the captions are generated by automatic image captioning systems, which allow us to obtain a rich description of images without any manual effort. Detecting such noise in the captions will require a joint understanding of the correlations between the text and image features.
- We propose a novel Image-Text Robust Retrieval Alignment Network (ITRANE) for a robust image-text retrieval approach. It utilizes a fine-grained alignment between the image-text modality as a regularization constraint to guide the robust learning process.
- Extensive experiments on two baseline datasets (COCO [74] and Flickr30k [132]) show that our proposed approach brings significant improvement in performance compared to multiple state-of-the-art image-text retrieval methods and noise-robust image-text retrieval methods in presence of noisy caption.

## 4.2 Related Work

**Image-text retrieval.** Matching between images and sentences is the key to image-text cross-modal retrieval. Most previous works exploited visual-semantic embedding to calculate the similarities between image and sentence features after embedding them into the joint embedding space [59, 35, 124, 140, 19, 34]. Coarse-grained matching approaches utilize multiple neural networks to compute a global feature and each network is used for a specific modality. For example, authors in [60] use a Convolutional Neural Network (CNN) and a Gated Recurrent Unit (GRU) to obtain the image and text features, while enforcing

the similarity of positive pairs larger than that of the negative ones. Inspired by common loss functions used in structured prediction, [35] focuses on hard negatives for training. The authors in [124] use soft gate for fusion to adaptively control the information flow for message passing between the modalities. With the emergence of transformers [120] in language understanding, recent works have explored the use of these architectures for vision and language tasks. Some such image-text [69, 87, 21, 63] or video-text[146] models rely on features extracted from each patch of images or video frames. A few other works, such as PixelBERT [50] and VirTex [28] operate directly on dense feature maps instead of relying on separate patches. All afore-mentioned works in image-text retrieval learn from clean data while our proposed retrieval system is capable of learning from a mixture of clean and noisy captions.

**Vision-and-Language Pre-training.** There is a recent growing interest in pre-training generic models to solve a variety of Vision-language problems, such as visual question answering, image-text retrieval and image captioning etc. Vision-language pre-training aims to improve performance of downstream vision and language tasks by pre-training the model on large-scale image-text pairs. For example, there are different architectures (e.g. two-stream models [56, 81, 80] vs. single-stream models [21, 63]), backbones (e.g. ConvNets [50] vs. Transformers [57]) etc. All these works aim to exploit the large-scale aligned image-text corpora to pre-train a powerful multi-modal model, which is then adapted to various downstream Vision-language tasks. Despite the use of simple rule-based filters, noise is still prevalent in large-scale texts crawled from web. However, the negative impact of the noise has shadowed by the performance gain obtained from scaling up the dataset.

**Learning with Noisy Data.** To handle the possible noisy annotations in the training data, a large number of methods have been proposed and almost all of them focus on the classification task. Noise-resistant methods use a wide variety of techniques such as robust losses [17], sample selection [40], regularization [79], and meta learning [65]. One approach to alleviate corrupted labels is based on noise correction strategy presented in [43, 115, 131, 41, 73, 121], aiming at removing samples with suspicious labels or correct their noisy labels to corresponding true ones. Authors in [73] propose a method to learn from noisy labels, by leveraging the knowledge learned from a small clean dataset and semantic knowledge graph to correct the noisy labels. The way these methods deal with noise require extra ground truth information indicating the presence of noisy labels. Providing ground-truth labels cannot always be guaranteed. Noise tolerant methods such as [55, 42, 133] focus on identifying true labeled samples from noisy training data. These methods to deal with noisy data are mostly applicable in other tasks, rather than multimodal learning.

**Image-Text Retrieval with Noisy Data.** Most prior works for noisy labels are specifically designed for unimodal scenarios. Works in [85, 47] investigated the problem of image-text retrieval with noisy labels and presented a two-step pre-processing method to obtain clean labels and feed them to cross-modal methods. To alleviate the influence of noisy samples, [47] reshaped the loss function to down-weight difficult samples and up-weight easy ones and thus focused training on clean samples. However, this method cannot be extended to the case where the caption itself is noisy. Another strategy known to overcome noisy caption is filtering before training. Work in [52, 112] exploits some degree of text filtering for this purpose. However, these pre-proprosseing steps are expensive, time-consuming

and ineffective at identifying mismatches between image and text features (see explanation earlier). In this paper, we propose a method to learn directly from noisy captions without any pre-processing step.

## 4.3 Methodology

This section provides the details regarding our proposed approach for the task of image-text retrieval with noisy captions using alignment model. Our goal is to train the model to output a similarity score between an input image $x$ and a textual description $c$. First, we define the problem statement (Section 4.3.1). Then we discuss the image-text retrieval model (Section 4.3.3) and the robust learning approach (Section 4.3.4) using an alignment model which associates detected tags/object labels with corresponding image regions and bridge the gap between them to ensure the related embeddings are appropriately contextualized in both modalities.

### 4.3.1 Problem Statement

Our objective is to learn a robust image-text retrieval system from a set of available images with noisy captions. Let $\mathcal{X}$ denote a set of $N$ images with noisy captions, i.e., $\mathcal{X} = \{(x_1, c_1), ..., (x_N, c_N)\}$, where $x_i$ is the $i^{th}$ image and $c_i$ is its corresponding noisy caption. The trained robust image-text retrieval system should be able to retrieve the relevant correct caption based on the given image, and similarly, retrieve the relevant image based on the given caption.

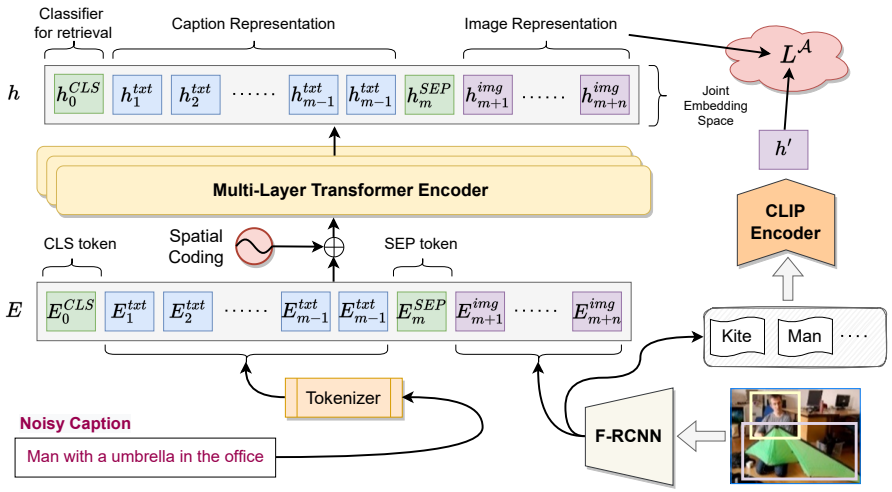



Figure 4.2: Our proposed framework consists of two main components: i) the base image-text retrieval system, ii) the Alignment Model, which associates detected tags/object labels with corresponding regions in an image. This compels the framework to learn correspondence embeddings similarly and bridge the gap between them to identify the noise in the caption.

### 4.3.2 Feature Extraction

A typical approach to attend to visual and textual features in the joint space is using transformer-based models. Employing the transformer architecture, we can simply operate on a concatenation of image and text, thereby learning the cross-modal contextualized features. Therefore, we deploy a vision-language transformer as our image and text encoder. We extract regions from input image, concatenate them with corresponding caption tokens and encode them as a sequence of embeddings, with an additional [CLS] token to represent the global image feature. For an input image-text pair, proper identification of the alignment requires fused vision-language representation to be accurate. Formally, assume we have a set $\mathcal{X}$ of $N$ images with noisy captions, i.e., $\mathcal{X} = \{(x_1, c_1), ..., (x_N, c_N)\}$, where $x_i$ is the $i^{th}$ image and $c_i$ is its corresponding noisy caption. For each image, an object detector is used to obtain $t$, where $t$ is a sequence of tags/labels.

Extracting the image and text features is the first and also the most crucial process in the retrieval system. Instead of extracting the global CNN feature from a pixel-level image, we focus on local regions and take advantage of bottom-up attention. We detect objects in each image utilizing a Faster R-CNN [99], which is pretrained on Visual Genome. For an image $x$, we extract features of $k$ regions $\{r'_1, r'_2, \ldots, r'_k\}$ using Faster R-CNN [99]. For each region, we concatenate 4 dimensional bounding box information predicted by Faster R-CNN to obtain position aware features. A linear layer is used to obtain projected region features $\{r_1, r_2, \ldots, r_k\}$, which has the same vector dimension as the word embeddings.

We use discrete tokens comprised of vocabulary words to represent captions. These tokens are the semantic units for processing the language. For each word in a caption, we use its representative token. Then the embedding of each word is the sum of its token-specific learned embedding [30] and encodings for position (token's index in the sequence). For each image-caption instance $(x, c)$, we combine the embeddings of image and text with embedings of some special tokens (classification token:`[CLS]`, separator token:`[SEP]`) sequentially following the format: '`[CLS] caption [SEP] image`'. This sequence is input to the image-text retrieval model for each image-caption pair.

### 4.3.3 Image-Text Retrieval Model

The image-text retrieval model consists of transformer encoder and a classification layer (see Fig. 4.2). As we can see in Fig. 4.2, the transformer encoder $\mathcal{E} : \mathcal{X} \rightarrow \mathcal{R}^d$ maps image-caption pair $(x, c) \in \mathcal{X}$ to a fused vision-language representation $\boldsymbol{h} \in \mathcal{R}^{d_h}$. Here, $d_h$ is the feature dimension of the fused vision-language representation $\boldsymbol{h}$ ($h = [h_0, h^{txt}, h^{Img}]$). The final representation of [CLS], $\boldsymbol{h_0}$ is used as the input to the classifier to predict whether

the given pair is aligned or not. We did not use ranking losses as we found that the binary classification loss works better, similarly as reported in [69]. A classifier $\mathcal{F} : \mathcal{R}^{d_h} \rightarrow \mathcal{R}$ is used to identify if the vision-text representation $\boldsymbol{h_0}$ corresponds to correct image-text pair. Combining the encoder $\mathcal{E}$ and classifier $\mathcal{F}$, we define the trainable image-text retrieval model $\mathcal{T}_\theta : \mathcal{X} \rightarrow \mathcal{R}$ parameterized by $\theta$. Here, $\mathcal{T}_\theta(x, c)$ outputs the probability of $(x, c)$ being the correct image-caption pair. **Our objective is to learn a robust $\mathcal{T}_\theta$ for image-text retrieval task.**

To train the image-text retrieval model $\mathcal{T}_\theta$, the retrieval task is formulated as a binary classification problem. For each given image-text pair $(x, c)$ with possibly noisy caption $c$, we formulate three negative/unaligned image-text pairs by i) sampling two random captions for the given image $(x, c_1^-)$, $(x, c_2^-)$, and ii) sampling a random image given the caption $(x^-, c)$. So for a given image-caption pair $(x, c)$ with possibly noisy caption $c$, we have a collection of image-text pairs $\boldsymbol{s}$ and a collection of labels $\boldsymbol{y}$ indicating if each of the pairs are aligned according to the provided supervision. We use the image-text retrieval model $\mathcal{T}_\theta$ to obtain the output probabilities of the image-caption pairs being aligned. Hence,

$$\boldsymbol{s} = \begin{bmatrix} (x, c) \\ (x, c_1^-) \\ (x, c_2^-) \\ (x^-, c) \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_{(x,c)} \\ y_{(x,c_1^-)} \\ y_{(x,c_2^-)} \\ y_{(x^-,c)} \end{bmatrix}, \quad \mathcal{T}_\theta(\boldsymbol{s}) = \begin{bmatrix} \mathcal{T}_\theta(x, c) \\ \mathcal{T}_\theta(x, c_1^-) \\ \mathcal{T}_\theta(x, c_2^-) \\ \mathcal{T}_\theta(x^-, c) \end{bmatrix}$$

To learn a image-text retrieval model that can predict the image-caption alignment cor-

rectly, cross-entropy loss, $l(\cdot)$, between predicted alignment scores and provided ground truth information is used. It aims to learn image-text multimodal representation that captures the fine-grained alignment between vision and language. As it is a binary classification task, the model uses a head (a linear layer) to predict whether an image-text pair is positive (matched) or negative (unmatched) given their multimodal feature. Here $L^{\mathcal{T}}\big(\boldsymbol{y}, \mathcal{T}_\theta(\boldsymbol{s})\big)$ is the cross-entropy loss computed for all four samples in $\boldsymbol{s}$, i.e.,

$$L^{\mathcal{T}}\big(\boldsymbol{y}, \mathcal{T}_\theta(\boldsymbol{s})\big) = \frac{1}{4}\sum_{j=1}^{4} l\big(\boldsymbol{y}[j], \mathcal{T}_\theta(\boldsymbol{s})[j]\big). \tag{4.1}$$

### 4.3.4 Aligment Model for Robust Learning

In this step, we carry out fine-grained fragment-level global-to-local matching between visual regions and textual words. With presence of noise in the caption, the biggest challenge is to design model architectures that can perform understanding-based tasks (e.g. image-text retrieval) while being robust to the noise. Towards this goal, in addition to using the image-text retrieval system (vision-language transformer) with image features and word embeddings of the noisy caption/description as input, we construct an alignment network. The alignment network progressively explore the local-to-local semantic alignments between image region and tags/object labels detected by Faster R-CNN [99]. Our alignment model assumes embedding of image regions and the corresponding tags/object labels detected by Faster R-CNN [99] should be close to each other in feature space. We calculate the image

region $h^{img}$and word $h^{txt}$ embedding distances using euclidean distance as follows:

$$||h^{img} - h^{txt}||_2^2. \tag{4.2}$$

Distances between image region $h^{img}$and word $h^{txt}$ embeddings give us key insight if the image region and corresponding word embeddings are learned close to each other during the training. Our alignment model which consists of a region encoder (transformer) and a label encoder (CLIP text encoder [95], projects the regional representation and label representation in the joint embedding space. Fixed embeddings generated by the CLIP text encoder compel our retrieval system to learn object region embeddings similar to corresponding correct labels, leading to robustness to noisy words in the caption. We want to define a scheme where corresponding region representation and label representation (positive pair) are brought together while irrelevant ones (negative pairs) are pushed away in feature space. To meet this criterion, we minimize the distance between positive pairs and maximize it between negative ones using triplet loss [108] as follows:

$$L^{\mathcal{A}}(h, h') = \frac{1}{k}\sum_{j=1}^{k}\left(||h_j^{img} - h'_j||_2^2 - ||h_j^{img} - h_j^{txt}||_2^2\right) \tag{4.3}$$

Here the positive input embedding $h'$ is selected based on F-RCNN label prediction for the image region. The negative input embedding $h^{txt}$ for each image region is chosen randomly from the nouns in the noisy caption excluding the corresponding label if there is any. $h^{img}$ is the image region embeddings as the anchor input ($h = [h^{txt}, h^{img}]$).

**Algorithm 4** Image-Text Robust Retrieval Alignment Network (ITRANE)

**Input:** Image $(x_i)$-text $(c_i)$ pair set $\mathcal{X}$ where the text is noisy.
**Output:** Robust image-text retrieval model parameters $\boldsymbol{\theta}$.
**Initialization:** Initialize image-text retrieval model parameters $\boldsymbol{\theta}$
**while** not converged **do**
  Fetch Mini-batch of image-text pairs $B^f = \{x_i^f, c_i^f\}_{i=1}^b \sim \mathcal{X}$
  Compute image-text allignment loss $L^{\mathcal{T}}(B^f; \theta)$ using Eq. (1)
  Compute alignment model triplet loss $L^{\mathcal{A}}(x^f; \theta)$ by Eq.(3)
  Update retrieval model parametrs $\theta \leftarrow \theta - \eta \nabla_\theta (L^{\mathcal{T}} + L^{\mathcal{A}})$
**Return** $\boldsymbol{\theta}$

**Final Objective.** The overall training strategy for our algorithm is presented in Algorithm 4. We jointly optimize two objectives including image-text retrieval loss $L^{\mathcal{T}}$ and alignment loss $L^{\mathcal{A}}$ using Equ.4.4. Our complete model utilizes a fine-grained alignment between the image-text modality as a regularization constraint to guide the robust learning process. To learn a regularized objective function, a constraint is also enforced on the model parameter $\theta$ with learning hyper-parameter of $\gamma$.

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} [L^{\mathcal{T}}\big(\boldsymbol{y}_i, \mathcal{T}_\theta(\boldsymbol{s}_i)\big) + L^{\mathcal{A}}(h_i, h_i')] + \gamma ||\boldsymbol{\theta}||_2^2. \tag{4.4}$$

## 4.4 Experiments

In this section, we experimentally evaluate the performance of our proposed approach. We first describe the experimental setups consisting dataset description, evaluation metrics, and the implementation details of the experiments. Finally, we report and analyze the results both quantitatively and qualitatively.

### 4.4.1 Experimental Setups

**Datasets.** We employ two commonly-used image-text pair datasets, COCO [74] and Flickr30k [132], to evaluate our proposed method. COCO is a large-scale object detection, segmentation, key-point detection, and captioning dataset. It contains 164K images split into training (83K), validation (41K) and test (41K) sets. Every image contains 5 text descriptions. Flickr30k is a standard benchmark for sentence-based image description. It contains 31783 images collected from Flickr, together with 5 reference sentences provided by human annotators. We split the dataset into 1000 test images, 1000 validation images, and 29783 training images.

**Generating Noisy Annotation.** To generate noisy captions, we exploit an image captioning model pretrained‘[83] on Conceptual Caption (CC) dataset [109]. Since the prediction/inference of captioning model can be inaccurate, feeding the images of COCO [74] and Flickr30k [132] datasets to the caption generator leads to a collection of noisy captions. We collect the noisy captions for the trainset of both COCO [74] and Flickr30k [132] dataset.

**Implementation Details.** Our training data consists of original images from COCO and Flickr30k trainset and the generated noisy captions. We evaluate on clean COCO and Flickr30k test set (i.e., both images and corresponding captions are from the original datasets). We employ a multi-layer transformer similar to the pretrained OSCAR [69] model as the baseline for the image-text retrieval task (ITRANE). The layers of the transformer encoder are frozen up to the last layer. The features for image regions are extracted using Faster-RCNN [99] pretrained on Visual Genome dataset [62]. To extract

the tags/label embeddings detected by Faster R-CNN, we use pretrained CLIP [95] text encoder. All the experiments are trained for 40 epochs with a batch size of 8. We use ADAM [58] optimizer with an initial learning rate of 0.00001. The decay factor in computing the loss moving average for curriculum learning is set to 0.90. The experiments are conducted on a NVIDIA 2080Ti GPU.

**Evaluation Metric.** We employ the Recall@K (R@K) metric for evaluating the retrieval performance for both image-to-text and text-to-image retrieval tasks. The R@K metric measures the percentage of queries for which our system is able to retrieve the correct item among the first K results. We consider K $\in \{1, 5, 10\}$. Please note that all the evaluations are done in testing sets with clean captions.

### 4.4.2 Result Analysis

The following experiments are conducted to evaluate the performance of our proposed approach:

- We analyze the significance of the problem, i.e., existing image-text retrieval model performance when trained with noisy text data.
- We compare the performance of proposed approach in presence of noisy caption during training with other image-text retrieval approaches.
- We analyze the advantage of using triplet loss with hard negative pairs and its effect on retrieval performance.
- We show qualitative examples from Faster R-CNN to observe its performance in detection of tags/labels, which helps identify noisy words in the caption.

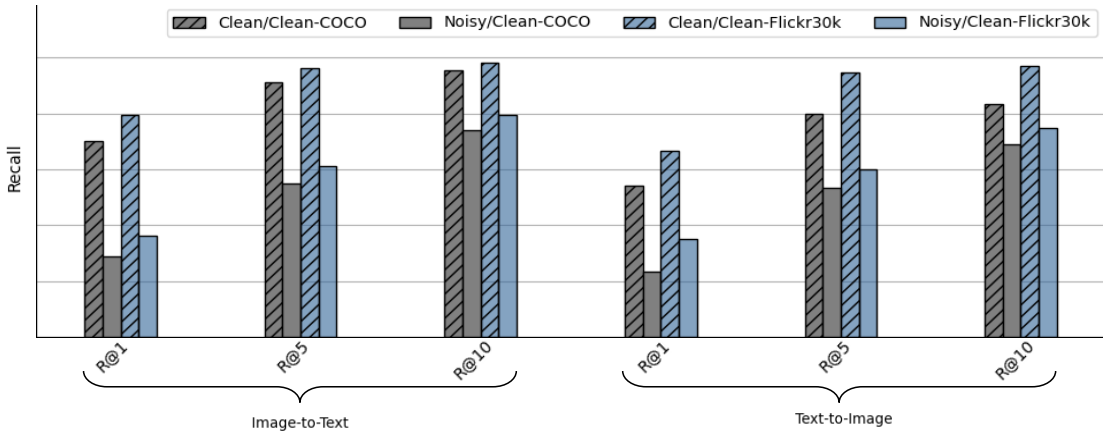



Figure 4.3: OSCAR [69] retrieval performance on COCO 5k [74] (grey) and Flickr30k [132] (blue) when train and test set both are clean vs when train set is noisy and test set is clean. We see performance drop in noisy training scenario.

- We visualize the learned semantic feature space of image-text pairs using alignment model.

- We show qualitative examples that verify the significance of the proposed approach.

**Significance of Robust Image-text Retrieval Approach.** Fig. 4.3 demonstrates the significance of designing noise robust image-text retrieval approach in presence of noisy caption. We consider OSCAR, a state-of-the-art image-text retrieval approach, fine-tuned for image-text retrieval task in COCO (gray) and Flickr30k (blue) dataset for two settings: i) the training images are associated with clean captions, and ii) the training images are associated with noisy/wrong captions generated by an automatic captioning process. OSCAR trained with noisy captions has a $15\% - 44\%$ performance drop compared to the same model trained with clean captions. We also observe a significant performance drop for other different approaches (VSE++, ReasoningNet). It demonstrates that image-text retrieval models trained with captions generated from automated captioning systems

are not robust to noise; thus require the development of noise-robust image-text retrieval systems.

**Robustness of ITRANE.** We evaluate the robustness of ITRANE for both image-to-text retrieval and text-to-image retrieval on COCO and Flickr30K datasets with noisy captions. Our baseline consists of three different categories of state-of-the-art image-text retrieval models: i) **VSE++** [35] (uses linear layer to project modality information in the joint embedding space), ii) **ReasoningNet** [87] (uses Transformer [120] to project the modality information), and iii) **OSCAR** [69] (Transformer based architecture that utilizes Masked-LM approach for pretraining). These models are trained with noisy captions for comparison. To show the efficacy of our proposed approach, we adapt both Transformer based image-text retrieval model **(ITRANE-T)** and BERT based pretrained image-text retrieval model with our proposed alignment model. Table 4.1 and Table 4.2 summarize the overall results for COCO and Flickr30K datasets respectively. For both datasets, we observe that conventional image-text retrieval models perform poorly when trained with noisy captions. ITRANE-T provides $4\% - 8\%$ improvement over its base approach ReasoningNet for both datasets over all the reported metrics. Our proposed ITRANE results in $4\% - 9\%$ absolute improvement over the best baseline (OSCAR) for both datasets. It validates that ITRANE is robust to noise present in captions.

We also compare the performance with existing noise-robust image-text retrieval approaches NCR [49] and DECL [94]. In the COCO dataset, ITRANE outperforms both NCR and DECL by a significant margin. In the Flickr30k dataset, ITRANE outperforms

Table 4.1: This table shows the performance comparison for the image-text retrieval task in the COCO dataset when models are trained with noisy captions. Proposed ITRANE achieves significant improvement over other approaches.

| **Method** | **Robustness** | **COCO 5k** | | | | | | **COCO 1k** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Image-to-Text** | | | **Text-to-Image** | | | **Image-to-Text** | | | **Text-to-Image** | | |
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VSE++ [35] | ✗ | 21.0 | 49.6 | 63.0 | 15.3 | 39.1 | 51.8 | 32.4 | 58.6 | 71.3 | 24.8 | 50.9 | 62.5 |
| ReasoningNet [87] | ✗ | 25.4 | 52.7 | 70.6 | 22.9 | 51.9 | 66.6 | 36.9 | 65.1 | 76.8 | 32.8 | 56.9 | 72.7 |
| OSCAR [69] | ✗ | 28.9 | 54.9 | 74.2 | 23.3 | 53.5 | 68.9 | 40.1 | 66.7 | 79.7 | 35.6 | 58.3 | 75.4 |
| NCR [49] | ✓ | 22.4 | 47.0 | 60.0 | 16.5 | 37.2 | 48.5 | 42.8 | 74.5 | 85.7 | 32.5 | 62.8 | 75.0 |
| DECL [94] | ✓ | 23.0 | 47.8 | 59.4 | 15.8 | 35.3 | 46.2 | 43.5 | 73.4 | 84.4 | 31.2 | 60.9 | 73.8 |
| **ITRANE-T** | ✓ | 29.3 | 59.4 | 77.8 | 26.5 | 56.1 | 72.8 | 42.1 | 70.9 | 82.7 | 37.6 | 61.5 | 77.1 |
| **ITRANE** | ✓ | **32.6** | **61.5** | **82.1** | **27.9** | **59.5** | **75.7** | **45.9** | **73.5** | **85.7** | **40.3** | **63.0** | **79.9** |

Table 4.2: This table reports the performance for the image-text retrieval task in the Flickr30K dataset when models are trained with noisy captions. Proposed ITRANE achieves significant improvement over other approaches.

| **Method** | **Image-to-Text** | | | **Text-to-Image** | | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VSE++ [35] | 25.2 | 54.9 | 66.4 | 15.9 | 44.4 | 53.7 |
| ReasoningNet [87] | 27.4 | 58.2 | 69.0 | 20.8 | 49.9 | 65.6 |
| OSCAR [69] | 33.5 | 60.3 | 79.5 | 29.1 | 59.9 | 74.8 |
| NCR [49] | **45.0** | **77.5** | **88.5** | 30.5 | 57.3 | 70.7 |
| DECL [94] | 19.9 | 45.5 | 58.0 | 12.0 | 28.2 | 37.9 |
| **ITRANE-T** | 31.1 | 66.3 | 75.1 | 24.9 | 55.1 | 70.8 |
| **ITRANE** | 37.4 | 69.3 | 88.3 | **33.9** | **65.6** | **81.2** |

Table 4.3: This table reports the performance for the image-text retrieval task averaged over all the datasets. The improved performance of ITRANE proves the generalizability of our proposed approach for robust image-text retrieval task.

| **Method** | **Image-to-Text** | | | **Text-to-Image** | | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VSE++ [35] | 26.2 | 54.4 | 66.9 | 18.7 | 44.8 | 56.0 |
| ReasoningNet [87] | 29.9 | 58.7 | 72.1 | 25.5 | 52.9 | 68.3 |
| OSCAR [69] | 34.2 | 60.6 | 77.8 | 29.3 | 57.2 | 73.0 |
| NCR [49] | 36.7 | 66.3 | 78.1 | 26.5 | 52.4 | 64.7 |
| DECL[94] | 28.8 | 55.6 | 67.3 | 19.7 | 41.5 | 52.6 |
| ITRANE-T | 34.2 | 65.5 | 78.5 | 29.7 | 57.6 | 73.6 |
| ITRANE | **38.6** | **68.1** | **85.4** | **34.0** | **62.7** | **78.9** |

DECL by a significant margin. Even though the performance of NCR is better compared to ITRANE for the image-to-text retrieval task in the Flickr30k dataset, ITRANE outperforms other methods significantly when averaged across both datasets (shown in Table 4.3), demonstrating its generalizability. ITRANE achieves up to 7% absolute improvement in image-to-text retrieval task and up to 14% improvement in text-to-image retrieval task over NCR across datasets.

**Random Negative Pairs vs Hard Negative Pairs.** To make the image-text

Table 4.4: Performance comparison of ITRANE model using negative vs hard negative input embeddings in alignment triplet loss. Different image-text retrieval approaches are trained on COCO 5k dataset. Using hard negative input embeddings provides the highest robustness compared to using random negative pairs and also compared to vanilla models (without alignment model).

| **Method** | $L_{\mathcal{A}}$ | **Image-to-Text** | | | **Text-to-Image** | | |
|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| ReasoningNet | - | 25.4 | 52.7 | 70.6 | 22.9 | 51.9 | 66.6 |
| ITRANE-T | N | 28.9 | 58.9 | 77.0 | 26.1 | 55.5 | 71.1 |
| **ITRANE-T** | HN | 29.3 | 59.4 | 77.8 | 26.5 | 56.1 | 72.8 |
| OSCAR | - | 28.9 | 54.9 | 74.2 | 23.3 | 53.5 | 68.9 |
| ITRANE | N | 32.2 | 61.0 | 81.5 | 27.5 | 59.0 | 75.1 |
| **ITRANE** | HN | **32.6** | **61.5** | **82.1** | **27.9** | **59.5** | **75.7** |

retrieval system robust to the noise in captions, our strategy is to define a scheme where corresponding region representation and label representation (positive pair) are learned closer to each other while irrelevant ones (negative pairs) are pushed away in feature space. To meet this criterion, we minimize the distance between positive pairs and maximize it between negative ones using triplet loss $L^{\mathcal{A}}$. The positive input embedding is selected based on Faster-RCNN label prediction for the image region. The negative input embedding for each image region is chosen randomly from the nouns in the noisy caption excluding the corresponding label if there is any. Inspired by hard negative mining, this strategy for selecting negative pairs increases the probability of incorporating hard negative input embeddings in triplet loss for the alignment model. Moreover, we can choose the negative input embedding randomly from the dataset dictionary rather than the corresponding noisy caption (random negative pairs instead of hard negative pairs). We analyze the robustness of our proposed approach based on the negative pairs selection strategy. In Table 4.4, we consider triplet loss $L^{\mathcal{A}}$ with both random negative (N) vs hard negative (HN) pairs.

We observe that when random negative pair is used, there is $4\% - 7\%$ improvement of performance compared to approaches that don't have any noise robust component. However, when hard negative pair is used, we obtain more improvement over the approaches that use random negative pairs in COCO dataset. Our intuition is that minimizing a loss function using hard negative mining helps learning corresponding region and word representation similarly while pushing away the closest irrelevant ones (noisy words in the caption). As a result, alignment model with hard negative pairs performs better compared to using random negative pairs and justifies the use of hard negative mining.

**Analysis of Visual and Textual Representation.** To check the performance of retrieval system in learning of distinguishable visual and textual representations, we visualize the learned semantic feature space of image-text pairs of the COCO test set on a 2D map using t-SNE [119] in Fig. 4.4. For each image region and word token, we pass it through the model, and use its last-layer output as features. The embedding of the image and the word corresponding to the same object share the same color. As we can see, the distance of the same object between two modalities is substantially low, i.e., most of them are perfectly aligned, as demonstrated by the overlapping regions. This shows learned representations correctly project object region and its label in the joint embedding space.

**Qualitative Analysis of Global-to-Local Matching.** Our method utilizes an alignment model which decomposes image-text matching into global-to-local level matching using pretrained object detectors. We equip our image-text retrieval system with the alignment model to associate detected tags/object labels with corresponding regions in an image. In Fig. 4.5, we qualitatively demonstrate Faster R-CNN tags/labels detection in two

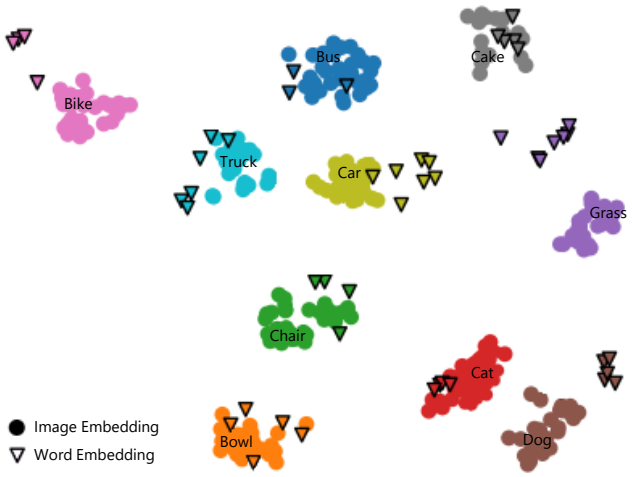



Figure 4.4: t-SNE visualization of region and word embeddings. We observe small distances between text and image embedding of the same object; most of them are perfectly aligned.

images from COCO [74] and Flickr30k [132] along with clean and generated noisy captions. In (a) and (b), correct object labels are detected for each image. For example consider Fig. 4.5 (b), the word "motorcycle" is replaced by "car" in the noisy caption. However, Faster R-CNN detects the label/tag "motorcycle" in the image precisely. Due to the use of the accurate and diverse object tags detected by Faster R-CNN (motorcycle, helmet, road), learning of corresponding image region representations is guided by correct label embeddings. These label embeddings are generated by pretrained CLIP encoder. This allows the network to identify noise in the caption by bridging the gap between image region and tags/labels embeddings, and compels it to learn more accurate image-text embeddings.

**Qualitative Results.** We qualitatively demonstrate some image-text pairs from COCO [74] and Flickr30k [132] along with generated noisy captions. Fig. 4.6 (i) elaborates the fact that generated captions consist of corrupted words which are semantically meaningful, while web-crawled captions used for training other image-text retrieval systems

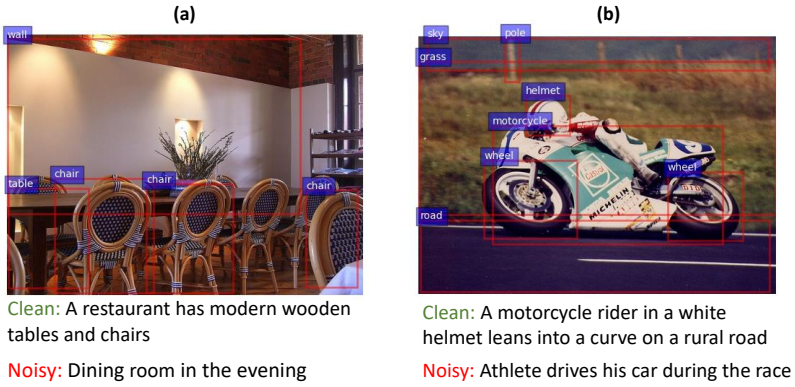



Figure 4.5: Faster R-CNN tags/labels detection in images from COCO [74] and Flickr30k [132] along with clean and generated noisy captions. In (a) and (b) images, correct object labels are detected which helps learning of corresponding image region representations even if they did not appear in noisy caption.

[52, 112] include nonsense errors which can be removed by filtering methods. In Fig. 4.6 (i), the captions in the first line of the description are the original texts, and the second rows include the generated noisy captions.

Fig. 4.6 (ii) shows top-2 ranked image-to-text and text-to-image retrieval results of two samples from COCO [74]. In the image-to-text retrieval case (top sample), captions in column (a) are retrieved by OSCAR [69]. Red captions show the wrong retrieval. As it is clear, OSCAR trained with noisy caption is not capable of retrieving any captions corresponding to the image. In column (b), our proposed framework retrieves the correct captions marked in green. In comparison to OSCAR [69], our method has promising performance as it retrieves two correct captions. The lower sample in Fig. 4.6 (ii) demonstrates the text-to-image case. Row (a) of images for the caption query is from OSCAR [69], and row (b) is for our proposed method. While OSCAR fails to retrieve the images corresponding to the query, our proposed method is able to retrieve results that connect the query and the corresponding image.

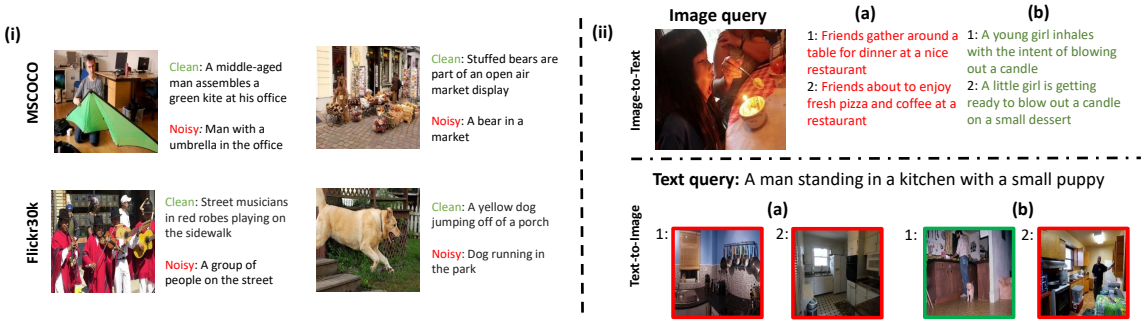



Figure 4.6: Qualitative retrieval results. (i) Examples of image-text pairs from COCO [74] and Flickr30k [132]. "Clean" refers to caption from the original dataset, and "Noisy" descriptions are generated noisy captions. (ii) The top-2 retrieved results are shown. Green denotes the ground-truth images or captions. Our model can capture the alignments between images and captions even though the training set consists of noisy captions.

## 4.5 Conclusions

The paper addresses an important practical problem, how to learn cross-modal retrieval methods from imperfect captions, thus allowing scalability to large-scale automatically captioned image datasets. Our proposed method utilizes an alignment model which decomposes image-text matching into global-to-local level matching using pretrained object detectors. Connection of detected tags/object labels with corresponding regions in an image compels the framework to learn correspondence embeddings similarly and bridge the gap between them to identify the noise in the caption. Experiments on two challenging datasets demonstrate our method has better performance in comparison to state-of-the-art baselines.

# Chapter 5

# Conclusions

## 5.1 Thesis Summary

In this thesis, we focus on robustness of vision-language systems to intentional and incidental adversity. We explore different data corruption categories including intentional and incidental errors. Data corruption refers to errors in data that occur during either an intentional process to alter data samples with the purpose of hiding information or an incidental data preparing step, including data annotation, which introduces unintended changes to the original data.

In Chapter 2, we address the problem of facial expression manipulation detection. We present a novel framework to exploit the underlying feature representations of facial expressions learned from expression recognition models to identify the manipulated features. Such approach achieves impressive performance in facial expression manipulation detection, while exhibiting robustness to other common facial corruptions like identity manipulation.

In Chapters 3 and 4, we propose robust image-text retrieval learning approaches. In Chapter 3, our proposed framework utilizes data-driven curriculum learning to up-weight the clean samples while discounting the noisy portion of the training data. Such sample weights, learned during training, compel the framework to learn from clean examples.

In Chapter 4, we equip our image-text retrieval system with an alignment model which bridges the gap between corresponding embeddings in different modalities to identify the noise in the caption. It aims to learn image-text multimodal representations that capture the appropriate fine-grained alignment between vision and language.

## 5.2 Future Research Directions

### 5.2.1 Noise Robust Retrieval Systems with Active Learning

Recent advancements in visual-textual retrieval and analysis tasks have been plagued significantly by the challenging and labor-intensive nature of annotating images/videos with text descriptions. Scaling up the dataset with noisy samples is one solution that recent works propose and we discuss it in Chapters 3 and 4. On the other hand, human interactive systems have attracted a lot of research interest in recent years, especially for image-text retrieval systems. Contrary to the early systems, which focused on fully automatic strategies, recent approaches have introduced human-computer interaction. To detect the noise in a dataset, one way is to use human feedback during training. To achieve the relevance feedback process, one can focus on the query concept updating. The aim of this strategy will be to refine the query according to the user labeling during training.

### 5.2.2 Robust Image-Text retrieval with Cooperative Distillation

Knowledge distillation (KD) is a model-independent knowledge transfer framework designed to deliver the knowledge extracted from a complex teacher model to a simple student model. Many existing studies have utilized KD to compress deep neural networks as well as to achieve stable performances by distilling the knowledge from teacher model. One extension to our robust image-text retrieval system would be cooperative distillation that combines clean curated datasets with noisy captions. Cooperative distillation framework trains a student model on a large noisy dataset. We also rely on a clean dataset to train a teacher model. We can explore leveraging the specific advantages of both types of datasets by training on a rich vocabulary and variety of scene contexts, while alleviating the noisy annotations.

### 5.2.3 Robust Video-Text Retrieval Systems

We have shown in Chapter 3 and 4 that robust image-text retrieval models help in improving time-consuming annotation process. One extension to our robust image-text retrieval system would be exploring the robust video-text retrieval systems. Specially, localizing video moments while there are noisy words in the caption is an interesting future direction of our work and can very helpful in many computer vision applications.

# Bibliography


[1] Fakeapp. https://www.fakeapp.com/.

[2] Google ai blog: Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[3] Faceswap. https://github.com/MarekKowalski/FaceSwap/, 2017.

[4] Deepfakes detection github. https://github.com/ondyari/FaceForensics/tree/master/dataset/DeepFakeDetection, 2019.

[5] Deepfakes github. https://github.com/deepfakes/faceswap, 2019.

[6] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, Dec 2018.

[7] Gary Anthes. Lifelong learning in artificial neural networks. *Communications of the ACM*, 62(6):13–15, 2019.

[8] Amir Bahmani, Arash Alavi, Thore Buergel, Sushil Upadhyayula, Qiwen Wang, Srinath Krishna Ananthakrishnan, Amir Alavi, Diego Celis, Dan Gillespie, Gregory Young, Ziye Xing, Minh Hoang Huynh Nguyen, Audrey Haque, Ankit Mathur, Josh Payne, Ghazal Mazaheri, Jason Kenichi Li, Pramod Kotipalli, Lisa Liao, Rajat Bhasin, Kexin Cha, Benjamin Rolnik, Alessandra Celli, Orit Dagan-Rosenfeld, Emily Higgs, Wenyu Zhou, Camille Lauren Berry, Katherine Grace Van Winkle, Kévin Contrepois, Utsab Ray, Keith Bettinger, Somalee Datta, Xiao Li, and Michael P. Snyder. A scalable, secure, and interoperable platform for deep data-driven health management. *Nature Communications*, 12(1):5757, Oct 2021.

[9] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, July 2019.

[10] Jawadul H. Bappy, Amit K. Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and B. S. Manjunath. Exploiting spatial structure for localizing manipulated image

regions. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] 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.

[12] Belhassen Bayar and Matthew C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, page 5–10, New York, NY, USA, 2016. Association for Computing Machinery.

[13] Tiziano Bianchi and Alessandro Piva. Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, 7:1003–1017, 2012.

[14] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[15] J. Bunk, J. H. Bappy, T. M. Mohammed, L. Nataraj, A. Flenner, B. S. Manjunath, S. Chandrasekaran, A. K. Roy-Chowdhury, and L. Peterson. Detection and localization of image forgeries using resampling features and deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1881–1889, July 2017.

[16] J. Bunk, J. H. Bappy, T. M. Mohammed, L. Nataraj, A. Flenner, B. S. Manjunath, S. Chandrasekaran, A. K. Roy-Chowdhury, and L. Peterson. Detection and localization of image forgeries using resampling features and deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1881–1889, July 2017.

[17] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, pages 961–970. PMLR, 2019.

[18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[19] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[20] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

[21] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[22] Giovanni Chierchia, Sara Parrilli, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. On the influence of denoising in prnu based forgery detection. In *Proceedings of the 2nd ACM Workshop on Multimedia in Forensics, Security and Intelligence*, MiFor '10, page 117–122, New York, NY, USA, 2010. Association for Computing Machinery.

[23] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, July 2017.

[24] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, page 159–164, New York, NY, USA, 2017. Association for Computing Machinery.

[25] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *CoRR*, abs/1812.02510, 2018.

[26] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection, 2018.

[27] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021.

[29] 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.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[31] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset, 2020.

[32] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and C. Canton-Ferrer. The deepfake detection challenge (dfdc) preview dataset. *ArXiv*, abs/1910.08854, 2019.

[33] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. *Towards Generalizable Deepfake Detection with Locality-Aware AutoEncoder*, page 325–334. Association for Computing Machinery, New York, NY, USA, 2020.

[34] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10792–10801, 2019.

[35] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[36] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7:1566–1577, 2012.

[37] Maximilian Fink, Ying Liu, Armin Engstle, and Stefan-Alexander Schneider. Deep learning-based multi-scale multi-object detection and classification for autonomous driving. In *Fahrerassistenzsysteme 2018*, pages 233–242. Springer, 2019.

[38] J. Fridrich and J. Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, June 2012.

[39] Chris Frith. Role of facial expressions in social interactions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364:3453–8, 12 2009.

[40] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 135–150, 2018.

[41] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. *arXiv preprint arXiv:1805.08193*, 2018.

[42] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018.

[43] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5138–5147, 2019.

[44] Behzad Hassani and Mohammad H. Mahoor. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, pages 790–795, 2017.

[45] 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 (CVPR)*, June 2016.

[46] Guosheng Hu, Yongxin Yang, Dong Yi, Josef Kittler, William Christmas, Stan Z Li, and Timothy Hospedales. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 142–150, 2015.

[47] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5403–5413, June 2021.

[48] 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.

[49] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29406–29419. Curran Associates, Inc., 2021.

[50] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[51] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

[52] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021.

[53] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[54] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, 2020.

[55] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR, 10–15 Jul 2018.

[56] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[57] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[58] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[59] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[60] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[61] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *ArXiv*, abs/1812.08685, 2018.

[62] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[63] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[64] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[65] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019.

[66] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009, 2020.

[67] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[68] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[69] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[70] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. *CoRR*, abs/1806.02877, 2018.

[71] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[72] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020.

[73] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017.

[74] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[75] Bo Liu and Chi-Man Pun. Deep fusion network for splicing forgery localization. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[76] Bo Liu and Chi-Man Pun. Deep fusion network for splicing forgery localization. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 237–251, Cham, 2019. Springer International Publishing.

[77] Junhao Liu, Min Yang, Chengming Li, and Ruifeng Xu. Improving cross-modal image-text retrieval with teacher-student learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3242–3253, 2021.

[78] Qingzhong Liu. Detection of misaligned cropping and recompression with the same quantization matrix and relevant forgery. In *Proceedings of the 3rd International ACM Workshop on Multimedia in Forensics and Intelligence*, MiFor '11, page 25–30, New York, NY, USA, 2011. Association for Computing Machinery.

[79] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[80] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[81] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[82] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. *arXiv preprint arXiv:1803.04376*, 2018.

[83] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. *arXiv preprint arXiv:1803.04376*, 2018.

[84] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pages 3355–3364. PMLR, 2018.

[85] Devraj Mandal and Soma Biswas. Cross-modal retrieval with noisy labels. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2326–2330, 2020.

[86] Ghazal Mazaheri, Niluthpol Chowdhury Mithun, Jawadul H. Bappy, and Amit K. Roy-Chowdhury. A skip connection architecture for localization of image manipulations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[87] Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. Transformer reasoning network for image-text matching and retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5222–5229. IEEE, 2021.

[88] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, Jan 2019.

[89] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress.

[90] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, 2019.

[91] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019.

[92] Yuichiro Nomura and Takio Kurita. Robust training of deep neural networks with noisy labels by graph label propagation. In *International Workshop on Frontiers of Computer Vision*, pages 281–293. Springer, 2021.

[93] Patrick Perez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, July 2003.

[94] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4948–4956, New York, NY, USA, 2022. Association for Computing Machinery.

[95] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[96] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1822–1830, 2017.

[97] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2017.

[98] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.

[99] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[100] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA, 2015. MIT Press.

[101] I.Michael Revina and W.R. Sam Emmanuel. A survey on human face expression recognition techniques. *Journal of King Saud University - Computer and Information Sciences*, 2018.

[102] F Rosenblatt. Th eperceptron: A perceiving and recognizing automaton. *Cornell University, Ithaca, NY, Project PARA, Cornell Aeronaunt Laboratory, Rep*, pages 85–460, 1957.

[103] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179, 2018.

[104] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[105] Seung-Jin Ryu, Matthias Kirchner, Min-Jeong Lee, and Heung-Kyu Lee. Rotation invariant localization of duplicated image regions based on zernike moments. *IEEE Transactions on Information Forensics and Security*, 8:1355–1370, 2013.

[106] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *CoRR*, abs/1905.00582, 2019.

[107] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, 2017.

[108] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[109] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[110] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[111] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks, 2020.

[112] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine

learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery.

[113] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[114] Wenyun Sun, Haitao Zhao, and Zhong Jin. A visual attention based roi detection method for facial expression recognition. *Neurocomputing*, 296:12–22, 2018.

[115] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018.

[116] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), July 2019.

[117] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *CoRR*, abs/1904.12356, 2019.

[118] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nieundefinedner. Face2face: Real-time face capture and reenactment of rgb videos. *Commun. ACM*, 62(1):96–104, December 2018.

[119] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[120] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[121] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017.

[122] Luisa Verdoliva and Paolo Bestagini. Multimedia forensics. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2701–2702, New York, NY, USA, 2019. Association for Computing Machinery.

[123] Limin Wang, Sheng Guo, Weilin Huang, and Yu Qiao. Places205-vggnet models for scene recognition. *arXiv preprint arXiv:1508.01667*, 2015.

[124] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5764–5773, 2019.

[125] Martin Wegrzyn, Maria Vogt, Berna Kireclioglu, Julia Schneider, and Johanna Kissler. Mapping the emotional face. how individual face parts contribute to successful emotion recognition. *PLOS ONE*, 12(5):1–15, 05 2017.

[126] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1480–1502, New York, NY, USA, 2017. Association for Computing Machinery.

[127] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR*, June 2019.

[128] C. Z. Yang, J. Ma, S. Wang, and A. W. C. Liew. Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Transactions on Information Forensics and Security*, 16:1841–1854, 2021.

[129] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[130] X. Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019.

[131] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025, 2019.

[132] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[133] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019.

[134] Zhenlong Yuan, Yongqiang Lu, Zhaoguo Wang, and Yibo Xue. Droid-sec: deep learning in android malware detection. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 371–372, 2014.

[135] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[136] Jichao Zhang, Yezhi Shu, Songhua Xu, Gongze Cao, Fan Zhong, Meng Liu, and Xueying Qin. Sparsely grouped multi-task generative adversarial networks for facial attribute manipulation. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 392–401, New York, NY, USA, 2018. Association for Computing Machinery.

[137] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. Context-aware attention network for image-text retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3533–3542, 2020.

[138] Zhongping Zhang, Yixuan Zhang, Zheng Zhou, and Jiebo Luo. Boundary-based image forgery detection by fast shallow cnn. *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2658–2663, 2018.

[139] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *Computer Vision – ECCV 2016*, pages 425–442. Springer International Publishing, 2016.

[140] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(2), May 2020.

[141] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, June 2016.

[142] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *CoRR*, abs/1412.6856, 2014.

[143] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839, 2017.

[144] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *CVPR*, June 2018.

[145] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[146] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020.