UNIVERSITY OF CALIFORNIA, MERCED

Deep Representation Learning for Multimodal Data Retrieval

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering Computer Science

by

Yuxin Tian

Committee in charge:

Professor Shawn Newsam, Chair
Professor Ming-Hsuan Yang
Professor Shijia Pan

Summer 2023

The dissertation of Yuxin Tian is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

(Professor Shawn Newsam, Chair)

(Professor Ming-Hsuan Yang)

(Professor Shijia Pan)

University of California, Merced

Summer 2023

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation and gratitude to my advisor, Professor Shawn Newsam, for his invaluable guidance, support, and encouragement throughout my Ph.D. journey. His generosity and mentorship have significantly alleviated the challenges of this journey. His profound expertise, unwavering support, and insightful advice have guided me during the most demanding periods. He has invested numerous hours in discussions, brainstorming sessions, and revisions of my work, consistently providing enlightening feedback and constructive criticism. Working with him has been an immense pleasure and honor, and I deeply appreciate his continuous guidance and assistance over the past five years.

I extend my gratitude to my lab colleagues, Yi Zhu, Xueqing Deng, Haolin Liang, Shrishail Baligar, Ruiqian Zhang, Jianan Chen, Akshay Bhatia. The time spent with you on campus has been filled with joy and learning. Your camaraderie and shared insights have enriched my Ph.D. experience immeasurably.

I am also incredibly grateful to many exceptional collaborators outside of UC Merced. My mentors Dalton Lunga, Kofi Boakye, Yunzhong He have had a significant impact on my research journey. Our close collaboration has not only enhanced my knowledge and understanding but also provided me with different perspectives on my work.

To my parents, I owe my sincerest gratitude. Your unwavering faith and steadfast support throughout this process have been a source of immense strength for me. Your belief in my abilities and your constant encouragement have made this journey possible. I cannot thank you enough for the sacrifices you have made and the love you have shown throughout my life and during my time as a doctoral student.

Finally, to all my friends and well-wishers who have supported me along this journey, your understanding, support, and encouragement have not gone unnoticed. I am forever grateful for your companionship and belief in my abilities.

VITA

2015	Bachelor in Biotechnology, South China Agricultural University, China
2018	Master of Medicine (Pharmacology), Zhejiang University, China
2023	Ph. D. in Electrical Engineering and Computer Science, University of California, Merced

PUBLICATIONS

Que2Engage: Semantic Retrieval for Relevant and Engaging Products at Facebook Marketplace, Yunzhong He*, Yuxin Tian*, Mengjiao Wang, Feier Chen, Licheng Yu, Maolong Tang, Congcong Chen, Ning Zhang, Bin Kuang and Arul Prakash, The ACM Web Conference (WWW), 2023. (*Equal contribution.)

Fashion Image Retrieval with Text Feedback by Additive Attention Compositional Learning, Yuxin Tian, Shawn Newsam, Kofi Boakye, *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023.

Image Search with Text Feedback by Additive Attention Compositional Learning, Yuxin Tian, Shawn Newsam, Kofi Boayke, *Bay Area Machine Learning Symposium (BayLearn)*, 2022.

Model Assumptions and Data Characteristics: impacts on Domain Adaptation in Building Segmentation, Philipe Dias, Yuxin Tian Shawn Newsam, Aristeidis Tsaris, Jacob Hinkle, Dalton Lunga, *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2022.

AutoAdapt: Automated Segmentation Network Search for Unsupervised Domain, Xueqing Deng, Yuxin Tian, and Shawn Newsam, CVPR Workshop on Neural Architecture Search: 1st lightweight NAS challenge and moving beyond (CVPRW), 2021.

Scale Aware Adaptation for Land-Cover Classification in Remote Sensing Imagery, Xueqing Deng, Yi Zhu, Yuxin Tian and Shawn Newsam, *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Generalizing Deep Models for Overhead Image Segmentation Through Getis-Ord Gi* Pooling, Xueqing Deng, Yuxin Tian, and Shawn Newsam, *International Conference on Geographic Information Science*, (GIScience), 2021.

Cross-Time and Orientation-Invariant Overhead Image Geolocalization Using Deep Local Features, Yuxin Tian, Xueqing Deng, Yi Zhu and Shawn Newsam, *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

ABSTRACT OF THE DISSERTATION

Deep Representation Learning for Multimodal Data Retrieval

by

Yuxin Tian
Doctor of Philosophy in Electrical Engineering Computer Science

University of California Merced, Summer 2023

Professor Shawn Newsam, Chair

This dissertation explores the potential of deep representation learning in the realm of semantic matching across multimodal data - an area of increasing relevance as digital information becomes progressively diverse. With a core emphasis on enhancing the effectiveness of retrieval systems, this research dives into the intricacies of two specific types of deep representations: invariant representations and multimodal representations.

Invariant representations boost the resilience of deep representations to variations inherent in data, catering to changes such as alterations in image content, the timestamp of image creation, and orientation. This dissertation delves into the development of robust invariant representations that persist despite temporal shifts in the data, highlighting their utility in a variety of real-world applications. The applicability of the invariant representations is further examined in overhead image geolocalization.

Concurrently, multimodal representations aim to establish semantic correspondences across different data modalities, including text, images, tabular data, and more. The study of multimodal representations facilitates many applications such as image and text interleaved search engines and recommendation systems. We propose a framework for learning composed image-text representations. This approach combines visual and textual modalities to enrich the search experience, facilitating image retrieval supplemented by textual feedback. Due to the complexity of the recommendation system, optimizing the retrieval model alone may not always lead to the better performance. Thus, we propose a multi-task learning approach for multimodal representation learning to address

this challenge, thereby fostering more accurate semantic matching.

By extensively exploring deep representation learning for retrieval tasks, this dissertation illustrates the substantial potential inherent in learning invariant and multimodal representations. As such, it not only advances current understanding and development in this rapidly evolving domain but also lays the groundwork for future research opportunities.

Chapter 1

Introduction

1.1 What's Retrieval?

In recent years, we've seen an explosion of content generation and sharing across various domains, including social media platforms, medical imaging, and so on. This surge of information has brought new challenges, particularly in searching through vast databases to retrieve specific content. Retrieval systems are the cornerstone of how we interact with the vast amounts of data in the digital world. From the web search engines that help us navigate the Internet to the recommendation algorithms that curate personalized content on various platforms, retrieval systems are deeply ingrained in our information consumption patterns.

The retrieval process involves comparing representations of a query to those of documents to identify, retrieve, and rank documents that may be relevant in response to a specific query. Both the query and documents can encompass various modalities, such as image, text, and tabular data. Over the past few decades, a wide range of techniques have been developed to enhance retrieval quality across diverse domains. These domains range from geospatial information systems to social media platforms, and search engines. Each of these domains has unique requirements and challenges, pushing the boundaries of retrieval methodologies and inspiring a rich variety of strategies to enhance retrieval effectiveness and efficiency.

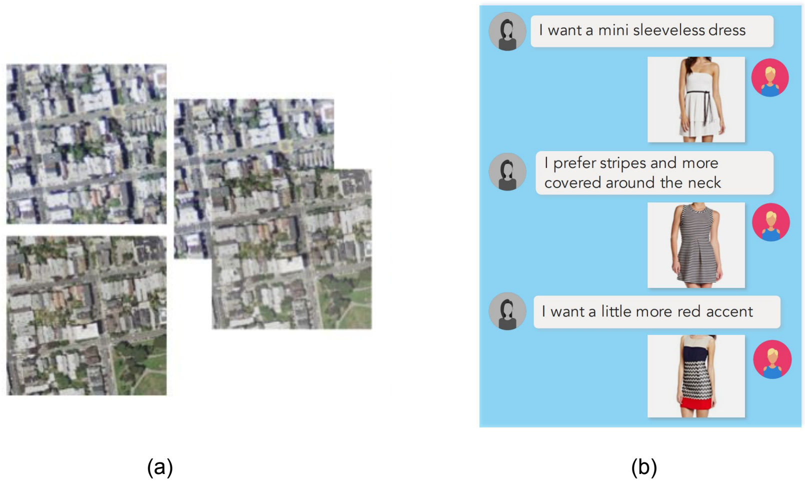
One can view the retrieval problem as learning a scoring function, denoted as $f : X \times Y \rightarrow \mathbb{R}$, which maps a pair of a query and a document, $(q, d) \in X \times Y$, to a score $f(q, d)$. The function is designed so that relevant (q, d) pairs have high scores, while irrelevant pairs score low. During the inference time, the system will return the documents with high scores based on the comparison with the search query. Beyond query-document retrieval, many real-world applications can be structured in this form. For instance, in geolocalization, q represents a query image, while d corresponds to similar images, but with known locations, from the same geographical location as the query image. In the advanced multimodal search engine, where image search is combined with text feedback, q represents a composite query of image and text, and d denotes the target image that aligns with the user's intent. In recommendation systems, q represents a user query, and d represents a potential product to recommend.

It is not always possible to retrieve identical documents from a large corpus for the given search query: the identical document may not exist in the document dataset, the intent behind a search can be vague and ambiguous, and the <query, target> may belong to different modalities. The task of accurately capturing search intent from a query and understanding the semantic representations of documents presents substantial challenges. As a result, various methods have been proposed to tackle the issue of semantic matching, moving beyond the traditional term matching methods or content-based image retrieval methods [47]. The goal of semantic matching is to address situations where the desired results are not exact matches of the query but still fulfill the users' search intent. Search queries can be formulated in various ways, including text queries, query images, or a combination of both, providing users with more convenient ways to express their needs.

1.2 Deep Representations Learning

Deep learning is crucial for "understanding" the content of raw data, where the focus lies on employing neural networks to extract suitable embeddings or representations from raw data. Fundamentally, embedding is a method of converting sparse vectors of IDs or pixel values into dense feature vectors, often referred to as semantic embedding due to its ability to capture semantics [47]. Learned embeddings are important for various downstream tasks, such as mortality prediction [127], bot detection [126], and retrieval. For the retrieval task, deep representations can be utilized to represent queries and documents for semantic matching. Specifically, in this thesis, we consider two types of deep representations for semantic matching in retrieval: invariant representations and multimodal representations.

Invariant representations enhance the resilience of deep representations to variations in data, such as changes in image content, image creation time, and orientation. Time-invariant features are characteristics of a model that remain unaffected by temporal changes in the input data. On the other hand, orientation-invariant features involve the model's capacity to identify and process input data irrespective of the orientation or rotation of objects or scenes in the data. These properties are particularly valuable in



I want a mini sleeveless dress

I prefer stripes and more covered around the neck

I want a little more red accent

(a)

(b)

Figure 1.1: Example scenarios where deep representation learning can play a role. (a) Perform geolocalization across time. (b) Advanced retrieval system with image and text. Image credit: FashionIQ dataset [39].

computer vision tasks, where objects may evolve over time or appear in various orientations within images. In Chapter 2, we present a method for cross-time and orientation-invariant overhead image geolocalization using semantic image matching.

Although significant progress has been made in terms of learning representations for vision or languages for retrieval [64, 33, 114], it is theoretically insufficient to model a complete set of human concepts using only unimodal data. However, establishing semantic correlations between different types of data is far from straightforward. For instance, understanding how a sentence in English corresponds to a specific image or video clip involves complex mappings that cannot be readily achieved through conventional methodologies. Due to the increasing popularity of social networks (e.g., Instagram, Twitter, Pinterest), the large amount of multimodal data containing both images and text on social websites has helped with the development of multimodal representation learning. Multimodal representations, bridging the heterogeneity gap among different modalities, play an indispensable role in the utilization of ubiquitous multimodal data. As a result of its powerful representation capabilities and the growing amount of multimodal data online, deep learning-based multimodal representation learning has garnered considerable attention in recent years. This dissertation demonstrates how to learn multimodal representations from images and their associated text modifications to retrieve the images with desired characteristics. I also introduce how to utilize image, text, and tabular data to learn representations for recommendation system.

1.3 Contributions

The contributions of the dissertation include:

- We have developed a method for large-scale overhead image geolocalization by comparing a query image to wide-area reference imagery with known locations. Our approach uses deep local features, allowing the query image to only overlap rather than align precisely with the reference imagery. We address the issues of images from different dates through cross-time geolocalization using time-invariant features, trained via a Siamese network. For differently oriented query and reference imagery, we have introduced an orientation normalization network. Through extensive experimentation, we demonstrate that our method outperforms existing state-of-the-art approaches.

- We introduce a novel solution that employs a multimodal transformer-based architecture for the fusion of image-text representations. This solution significantly enhances performance on several large fashion datasets, setting new standards in the challenging field of image search with text feedback.

- We present Que2Engage, a multimodal embedding-based retrieval system. This system integrates contextual signals as a unique modality in its transformer fusion backbone, striking a balance between relevance and engagement in a real-world recommendation system. Our proposed multimodal retrieval model effectively incorporates images, text, and contextual signals, as our results demonstrate.

1.4 Dissertation Overview

The remainder of this dissertation is organized as follows: Chapter 2 presents the method for learning invariant representations, fundamental to overhead image geolocalization. In Chapter 3, we delve into the learning framework for composed image-text representations derived from fashion imagery. This framework paves the way for an advanced search experience—image retrieval complemented with text feedback. Chapter 4 outlines the process of learning and optimizing multimodal representations during the retrieval phase within a complex recommendation system. Finally, in Chapter 5, we conclude the dissertation by discussing potential future research opportunities and possible extensions of the work presented.

Chapter 2

Cross-Time and Orientation-Invariant Overhead Image Geolocalization

2.1 Overview

Overhead image geolocalization is becoming increasingly important due to the growing collection of drone imagery without location information. In this chapter, we perform large-scale overhead image geolocalization by matching a query image to wide-area reference imagery with known location. We use deep local features so that the query image need not align with but only overlap the tiled reference imagery. We further address two key challenges. For when the query and reference imagery are from different dates, we perform cross-time geolocalization using time invariant features learned using a Siamese network. For when the query and reference imagery are oriented differently, we introduce an orientation normalization network. We demonstrate our contributions on two new high-resolution overhead image datasets. Our method significantly outperforms strong baselines on cross-time geolocalization and is shown to exhibit promising orientation invariance.

2.2 Introduction

While there has been a fair amount of work on locating ground level imagery [5, 42, 71, 90, 121, 135], there has been little work on the overhead case [29]. However, we believe this is an increasingly important problem due to the ease with which anyone can capture overhead imagery using drones and share it online. While location information typically accompanies traditional overhead imagery, such as from satellite and aerial platforms, location information is often missing or unreliable for drone imagery. It might become lost as the imagery is distributed or deliberately obscured. Our focus on overhead imagery geolocalization is thus timely and important.

This chapter focuses on the problem of geolocating overhead imagery captured from satellite, aerial, or drone platforms. By geolocating we mean assigning geographic coordinates such as latitude and longitude values. We allow the “search region” to be large and so this is a difficult problem. As shown in Figure 2.1, we formulate the problem as matching a query image to tiled wide-area reference imagery with known location. We address two fundamental challenges: the query and reference imagery 1) might have

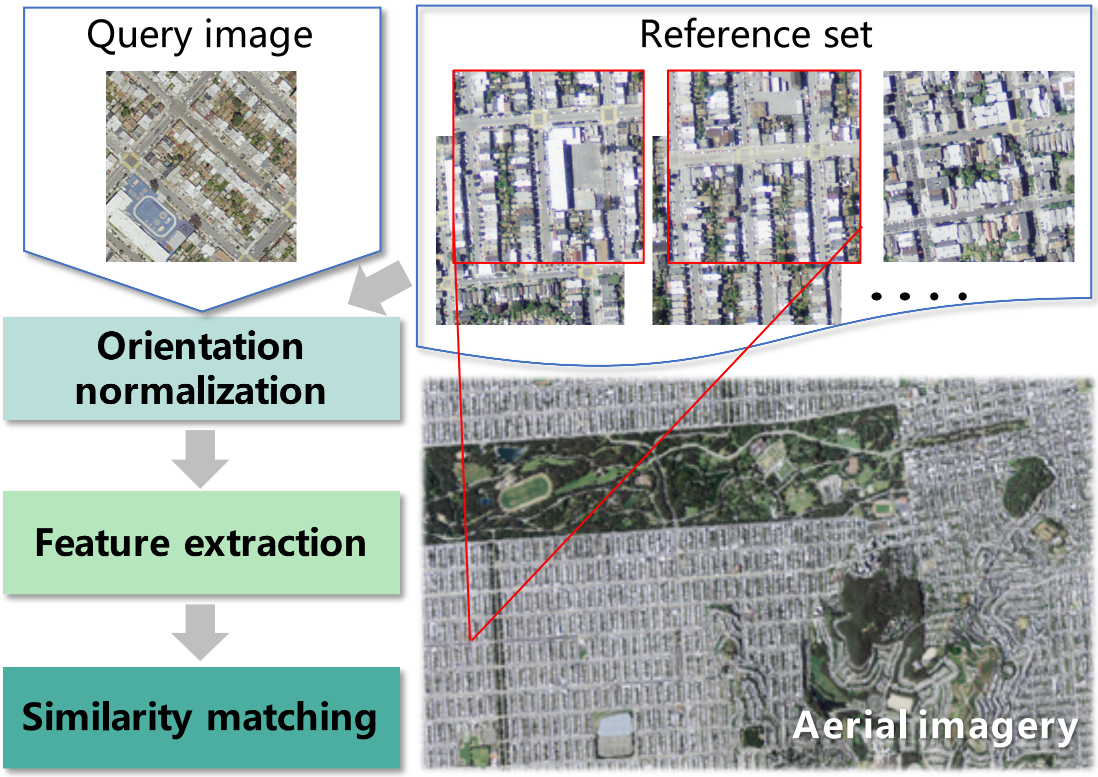


Figure 2.1: We perform overhead image geolocalization by matching a query image to reference imagery with known location. We address two fundamental challenges: cross-time matching and orientation-invariant matching. Compare the query image with the reference set above.

been taken at difference times, and 2) might be oriented differently.

We exploit recent advances in deep learning, particularly convolutional neural networks (CNNs), to perform the image matching. We show that, as expected, global image features extracted using the fully connected (*fc*) layers are not appropriate due to query-reference tile misalignment and so we instead derive local features from the locality preserving feature maps of the convolutional (*conv*) layers. We show these deep local features significantly outperform traditional local features, such as Scale Invariant Feature Transform (SIFT) features [75], when the query and reference images are from different dates. We next develop a Siamese network to explicitly learn time-invariant features to make our approach even more robust to changes in season, illumination, and sensors, and to changes in what is on the ground [92]. Finally, we tackle the real but challenging problem of when the query and reference imagery are oriented differently (*e. g.* , not both pointing north). For this, we develop an Orientation Normalization Network (ONN) that rotates the query and reference imagery to the same canonical orientation. We demonstrate our methods on two new high-resolution overhead image datasets.

The key and novel contributions of our work include:

- We perform large-scale overhead geolocalization via image matching using learned time-invariant deep local features.
- We propose an Orientation Normalization Network to account for when the query and reference imagery are oriented differently.
- We introduce two high-resolution overhead image datasets which will be made publicly available for other researchers.

2.3 Related Work

2.3.1 Image geolocalization.

Estimating the geographic location of an image has been of interest to the computer vision community for some time [5, 67, 69, 72, 95, 96, 97, 111, 122]. However, the focus has been mostly on geolocating *ground level imagery* which is a related but different problem than ours, and has a different set of challenges. Ours is one of the first works to focus on overhead image geolocalization.

Ground level imagery has been geolocated by matching it to maps in geographic information systems (GIS) [17], to other ground level imagery with known location [5, 42, 71, 90, 121, 135], to overhead imagery [7, 46, 69, 72, 104, 111, 122, 136], or to combinations of this reference data [67].

Geolocating a ground level query image by matching it to ground level reference imagery is limited to regions where reference imagery is available such as in urban areas [89, 88, 108] or along roads. It also typically assumes the query and reference images are both oriented with the sky at the top. We instead can geolocate overhead imagery from anywhere and which might be oriented differently from our reference imagery.

The fundamental challenge to geolocating a ground level query image by matching it to overhead imagery is the difference in perspective and so most of the work on this problem focuses on cross-view matching [69, 72, 97]. For example, Shi *et al.* propose a novel Cross-View Feature Transport (CVFT) layer to facilitate feature alignment between ground and aerial domains [97]. In contrast, our query and reference imagery are taken from the same viewpoint and so we face a different set of challenges.

We also expect to be able to geolocate overhead imagery more accurately than ground level imagery.

We know of only one other work on geolocating overhead imagery [29]. It also uses an image matching framework but uses traditional local features and does not address the cross-time or orientation-invariant cases. We include it as one of our baselines.

2.3.2 Orientation alignment.

The concept of orientation is very different for overhead images than for images taken at ground level. Most ground level images have a canonical orientation [35]. Street view images and the like typically have the ground at the bottom and sky at the top. Most objects have a canonical orientation when viewed from the side which then dictates the canonical orientation of the image [111]. In fact, researchers have exploited this fact to learn better representations in an unsupervised manner [37]. In contrast, overhead imagery typically does not have a canonical orientation. While most overhead imagery is oriented so that north points up, this has nothing to do with the content of the image and cannot be derived from it in the general case. There has been work on classifying rotation agnostic images [35] in the ImageNet dataset [27] by splitting the image representation into rotation related and unrelated parts. This, however, produces global features which are not appropriate for our problem.

CNNs are inherently limited in their ability to model geometric transformations due to the fixed geometric structure of their constituent modules [26]. Modules have been proposed that enable spatial manipulation, including rotation, of data within the networks [26, 49, 123]. We utilize one such module, Spatial Transformer Networks [49], in our Orientation Normalization Network below.

2.3.3 Image retrieval.

Our matching framework has many similarities with image retrieval methods. We distinguish it, though, from the following two main image retrieval paradigms. Similarity-based image retrieval methods seek to retrieve similar images and not necessarily images of the same scene. Retrieving similar images is not sufficient to geolocate overhead imagery since many locations might look very similar from above. Indeed, our results show that even when our matching framework fails, it still retrieves similar images. We need our matching to be more discriminating (yet still allow for differences due to time and orientation). Image retrieval has been used to geolocate ground-level imagery by matching it against ground-level images of the same scene. This is the approach taken by Radenovic *et al.* [90] using a method called fine-tuning image retrieval (FITR). These approaches tend to use global features though, which, as we will demonstrate, are not effective for our problem. We include FITR as one of our baselines.

2.4 Methodology

We formulate overhead image geolocalization as an image matching problem in which a query image is matched to wide-area reference imagery with known location. We assume the search area is covered by one contiguous reference image even though, in reality, it will be a registered mosaic of large but individually acquired images. In order to localize the matching, we partition the contiguous reference image into tiles the same size as the query (this also enables easy parallelization). Our problem thus reduces to finding a good representation $F(.)$ for overhead imagery so that given a query image q, we are able to find at least one spatially overlapping tile r from reference set R by computing the distance between $F(q)$ and $F(r)$.

Several things make this challenging. First, the query image is randomly located and thus not aligned with any of the reference tiles. The query image overlaps several of the reference tiles but by varying amounts and so we have to be able to perform matching

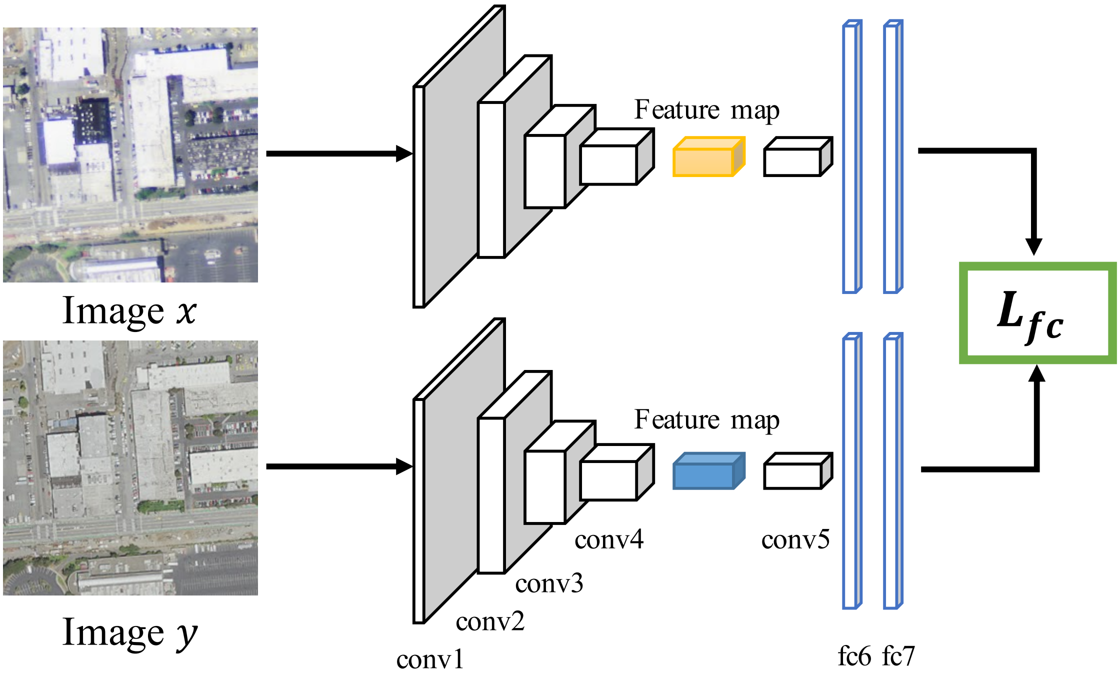


Figure 2.2: Siamese network for learning features for cross-time matching. The positive training samples are co-located images from different times.

based on this varying overlap. Second, the query and reference imagery might have been taken at different times, for example, when geolocating current drone imagery using archived satellite imagery. And, third, they might not have the same orientation. We describe our novel technical contributions to overcome these challenges in the following.

2.4.1 Deep Local Features

CNNs have proven effective at mapping images to powerful and often semantically rich feature vectors [62, 141]. Most work utilizes global features extracted from the fully connected layers including the work mentioned above on geolocating ground level imagery by matching against ground or aerial images [5, 69, 72, 90, 97, 111]. However, since our query and reference tiles only overlap, using global features to perform the matching is unlikely to be effective. Our results below demonstrate this.

We instead extract deep local features from the *conv* layers since locality is preserved in the feature maps. We split these feature maps along the channel dimension to produce a set of deep local features. Specifically, given an image x, we apply a trained CNN to compute a $conv$ layer output $F(x)$ of size $H \times W \times C$, where $H \times W$ are the spatial dimensions of the feature map and C is the number of channels. $F(x)$ is then split into a set of $H \cdot W$ vectors of length C. We denote these features as $\mathbf{p}_x^i$ where x is the image and i is the feature number which is in the range $(1, H \cdot W)$. Each image x, either query or reference, is thus represented by the set of deep local features $S_x = \{\mathbf{p}_x^i\}_{i=1}^{H \cdot W}$.

Matching between a query image and a set of reference tiles is then performed by finding, for each of the query's features, the nearest neighbor in feature space among all the features of the all reference tiles. Each nearest neighbor match votes for a reference tile and the votes are accumulated over all the query image's features to rank the reference tiles. Specifically, given a query image q with local features $\mathbf{p}_q^i$ and a set of reference tiles $r \in R$, each with local features $\mathbf{p}_r^i$, for each $\mathbf{p}_q^i$, we use the Euclidean distance to find the nearest neighbor:

$$\mathbf{p}_r^j = \arg \min_{r \in R, j=1,\ldots,H \cdot W} \|\mathbf{p}_q^i - \mathbf{p}_r^j\|_2. \tag{2.1}$$

This will result in a vote for reference tile r. We then rank the reference tiles in order of decreasing votes and pick the *top one* as the match for query image q. That is, we use only the best match among all the reference tiles to geolocate the query tile even though it overlaps multiple reference tiles. (See Figure 2.6.)

We first investigate deep local features extracted using a VGG16 network [101] trained on the ImageNet dataset [27]. These features are not specific to overhead image matching nor are they invariant to potential time differences between the query and reference images. One of our key technical contributions therefore is a Siamese network which learns improved deep local features specific to overhead imagery and for cross-time matching.

2.4.2 Siamese Network for Cross-Time Matching

Our proposed Siamese network is shown in Figure 2.2. It consists of two embedding CNNs that share weights. During training, the network is presented with either a pair of images from the same geographic location but taken at different times (positive examples) or a pair of images from different locations (negative examples). Positive

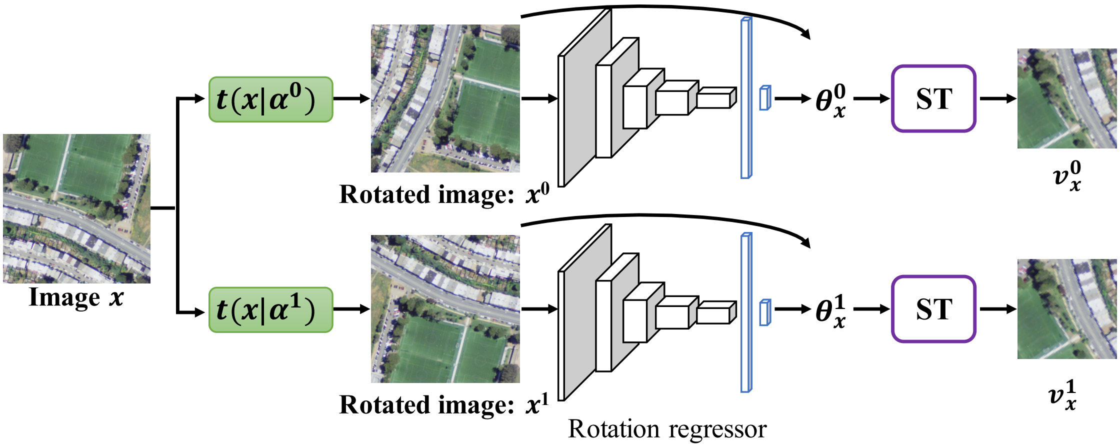


Figure 2.3: Orientation normalization network (ONN) which learns a rotation regressor to transform differently oriented images of the same location to the same orientation. ST is a spatial transformer layer.

examples are shown in Figure 2.5. The goal of the Siamese network is to learn a feature representation (non-linear embedding) $g(.)$ such that images from different locations are far apart in feature space while images from the same location are close *even if they are from different times*. This is done by training the network to minimize a contrastive loss [40]

$$L_{fc} = \frac{1}{2} l D^2 + \frac{1}{2}(1-l) \max\left(0, \left(m - D^2\right)\right), \tag{2.2}$$

where $l \in \{0, 1\}$ is the label indicating whether the input pair x, y is from the same location ($l = 1$) or not ($l = 0$), D^2 is the squared distance between $g(x)$ and $g(y)$, and m is the margin parameter that omits the penalty if the distance between images from different locations is too large.

Structurally, our Siamese network consists of two pre-trained VGG16 networks which we modify for fine-tuning on our overhead imagery. We remove the last fully connected layer $fc8$ and use the 4096-dim feature from $fc7$ to compute the Euclidean distance between $g(x)$ and $g(y)$. We investigate deep local features extracted from conv1 to conv4 of the trained embedding network in our experiments.

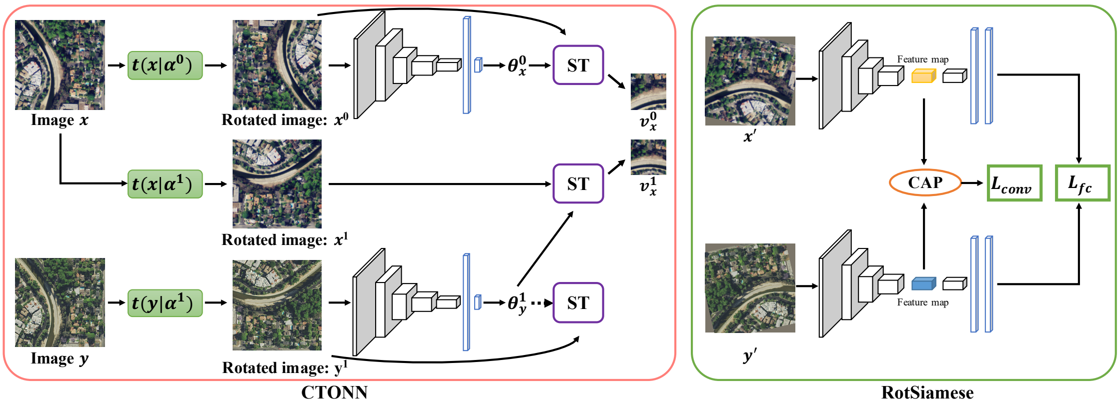


Figure 2.4: Our cross-time orientation normalization network (CTONN) learns a regressor that can orient images at different orientations and from different times to the same canonical orientation. Our RotSiamese network provides rotation invariance to deal with the noisy output of the CTONN at inference time.

2.4.3 Orientation Normalization Network

A fundamental challenge to performing overhead image geolocalization through image matching is that the query and reference imagery typically do not have the same orientation. While the reference imagery is usually oriented northwards, the orientation of the query image is generally arbitrary and unknown. Further, unlike ground level imagery, which has a standard orientation (such as the sky is up) that can be estimated and exploited by works like [37], there is no such standard orientation that can be estimated from overhead imagery.

Therefore, instead of trying to reorient the query image to a standard orientation so that it matches the reference imagery, we instead reorient both the query and reference to the same, potentially arbitrary direction. We seek a framework that can estimate such scene-specific canonical orientations.

Figure 2.3 shows our framework for learning a network that can be used to normalize the orientation of overhead images. This framework takes as input differently rotated versions of an overhead image and learns a rotation regressor that aligns the images. Specifically, we define a set of K discrete rotation transformations $T = t(.|\alpha^k)_{k=1}^K$, where $t(.|\alpha^k)$ is the operator that applies to image x the rotation transformation with angle α^k that yields the rotated image $x^k = t(x|\alpha^k)$. The α^k are evenly sampled from 0° to $360°$ depending on K. We investigate the choice of K in the experiments.

The goal of the rotation regressor in Figure 2.3 is to predict angles θ_x^0 and θ_x^1 such that when input image x^0 is rotated by θ_x^0, it has the same orientation as input image x^1 rotated by θ_x^1. If image x^0 was derived by rotating x by α^0 and image x^1 was derived by rotating x by α^1 , then the rotation regressor can be learned by minimizing the loss function L_θ

$$L_\theta = \left|\left(\alpha^0 + \theta_x^0\right) - \left(\alpha^1 + \theta_x^1\right)\right|, \tag{2.3}$$

where θ_x^k is the predicted angle for the rotated image x^k. (Note that the rotation regressor has no knowledge of α^0 and α^1.)

However, this objective alone leads to a trivial solution which predicts $\theta = 0$ regardless of the input. So, we modify the network to also compare the normalized images during training, that is the similarity of image x^0 rotated by θ_x^0 and image x^1 rotated by θ_x^1. We do this by inserting a spatial transformer (ST) layer [49] to produce images v_x^0 and v_x^1 (see Figure 2.3). Here, $v_x^k = ST\left[x^k|(\theta_x^k, rr)\right]$, where v_x^k denotes the transformed image whose input is x^k with the predicted rotation angle θ_x^k, and rr denotes the reduced ratio to crop the center of the rotated image in order to avoid introducing blank regions in the corners.

The rotation regressor is then learned by minimizing the joint loss function

$$L = \left|\left(\alpha^0 + \theta_x^0\right) - \left(\alpha^1 + \theta_x^1\right)\right| + \lambda_v \left|v_x^0 - v_x^1\right|, \tag{2.4}$$

which includes the L1 loss between images v_x^0 and v_x^1. The weighting parameter λ_v is chosen empirically.

We implement the rotation regressor using a VGG16 convolutional backbone followed by a two-layer regressor module to produce the angle θ. A scaled Tanh activation layer is appended to the regression model to constrain θ to meaningful values.

2.4.4 Cross-Time Orientation-Invariant Matching

We are only able to train our Orientation Normalization Network (ONN) using differently oriented images from the same time due to the sensitivity of the L1 loss to time-related differences. The learned rotation regressor is effective for normalizing images

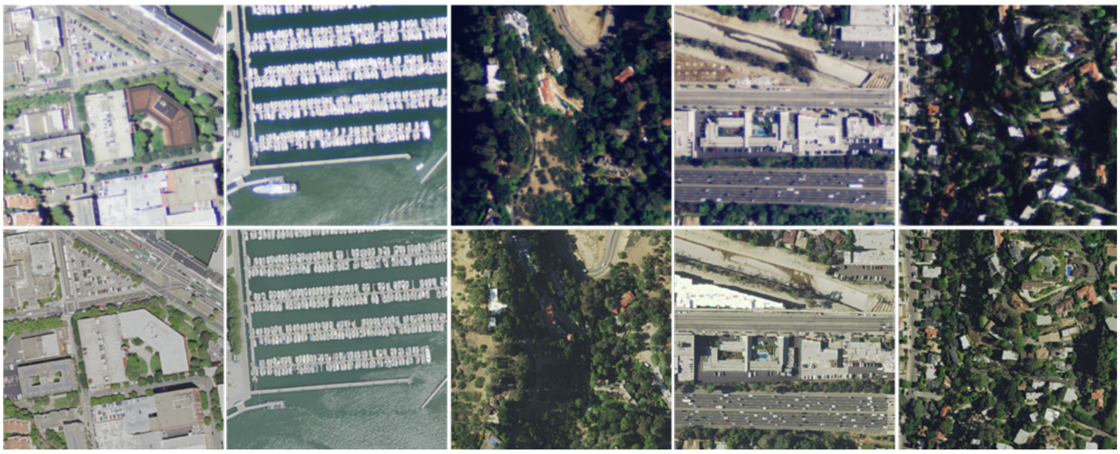


Figure 2.5: Co-located cross-time training pairs. Top: 2012. Bottom: 2014.

from the same time but has difficulties with images from different times. We therefore develop the cross-time ONN (CTONN) framework shown in Figure 2.4 left. This network now takes co-located image pairs x and y from different times and separately applies the random rotations α^0 and α^1 to produce x^0 and y^1. The rotation regressor is applied to predict θ_x^0 and θ_y^1 from these images. The spatial transformer module now applies rotation θ_y^1 to x^1 instead of y^1 to produce v_x^1, where x^1 is x rotated by θ_y^1 (the amount y was rotated to get y^1). v_x^0 and v_x^1 are now from the same year and can be compared using L1. The loss function for training CTONN becomes

$$L = \| \left(\alpha^0 + \theta_x^0\right) - \left(\alpha^1 + \theta_y^1\right) \| + \lambda_v \left| v_x^0 - v_x^1 \right| . \tag{2.5}$$

The proposed CTONN results in a rotation regressor that is more effective for normalizing the orientations of images from different times. However, the normalized images are still not aligned well enough for our cross-time feature extractor, which was trained using images with the same orientation. We therefore need to make our cross-time feature extractor more robust to these slight misalignments.

We develop the second Siamese network shown in Figure 2.4 right to make our feature embedding network more orientation invariant. We refer to this network as Rot-Siamese (RotSia for short). This network learns to extract deep local features that are more orientation invariant through 1) the addition of another loss term, and through 2) data augmentation. Specifically, the input images (same location different time) are separately rotated by small random angles sampled from a limited range $(-\phi, \phi)$ before being fed into the embedding network. This data augmentation alone does not result in improved orientation invariance as the loss function computed on the global features is not sensitive to slight differences in orientation. We therefore modify the loss to also compare the convolutional feature maps. We perform average pooling along the channel dimension (CAP) of the $conv$ layer that we use for deep feature extraction and compare these averages.

Specifically, given two rotated images x' and y', feature maps $F(x')$ and $F(y')$ are extracted from the $conv$ layer. The pooled feature maps $CAP(F(x'))$ and $CAP(F(y'))$ are then flattened and compared using the Euclidean distance:

$$D^2_{conv} = \|CAP(F(x')) - CAP(F(y'))\|. \tag{2.6}$$

This is then incorporated into a contrastive loss L_{conv}

$$L_{conv} = \frac{1}{2} l D^2_{conv} + \frac{1}{2}(1 - l) \max\left(0, \left(m - D^2_{conv}\right)\right). \tag{2.7}$$

Finally, the overall objective of the RotSiamese network is the weighted sum of this loss and the original one

$$L = L_{fc} + \lambda_c L_{conv}. \tag{2.8}$$

The weighting parameter λ_c is chosen empirically. Note that the CTONN and RotSiamese networks are trained separately but training them together in an end-to-end manner could be future work.

2.4.5 Geolocalization pipeline

Again, we perform geolocalization by matching the features of the query image to the features of the reference tiles and pick the top match through voting. When the query and reference tiles are from different times and have different orientations, we first perform orientation normalization on each tile separately using our trained CTONN and then extract deep local features using the feature embedding from the trained RotSiamese network. Figure 2.1 illustrates this pipeline. Note that the features can be pre-computed offline for all the reference tiles.

Dataset	2012		2014	
	SF	LA	SF	LA
Query	800	900	800	900
Reference	5569	6525	5569	6525

Table 2.1: The number of 256×256 pixel tiles in our dataset.

2.5 Experiments

In this section, we first introduce two new high-resolution overhead image datasets and describe our implementation details. We then we demonstrate our results on the cross-time, orientation-invariant overhead image geolocalization problem with comparison to strong baselines.

2.5.1 Dataset

We use high-resolution aerial imagery from the National Agriculture Imagery Program (NAIP) for our experiments. The images have a ground sample distance (GSD) of one meter (spatial resolution is 1m/pixel) and measure approximately $6k \times 7k$ pixels. We download eight pairs of spatially contiguous NAIP images from the San Francisco area and nine pairs from Los Angeles area. Each pair of images consists of co-located images but taken at different times, one in 2012 and the other in 2014. These pairs thus form our cross-time dataset. The reference datasets are constructed by partitioning the NAIP images into non-overlapping tiles measuring 256×256 pixels. The query images are not aligned with these reference tiles but are randomly extracted from the NAIP images and also measure 256×256 pixels. Table 2.1 summarizes the dataset.

During training, the Siamese networks and the cross-time ONN require pairs of co-located images from different times. We thus construct a training set of $12k$ pairs for SF and $13.5k$ pairs for LA. Examples of training pairs are shown in Figure 2.5. The negative training samples are cross-time images from different pairs to ensure they are not co-located.

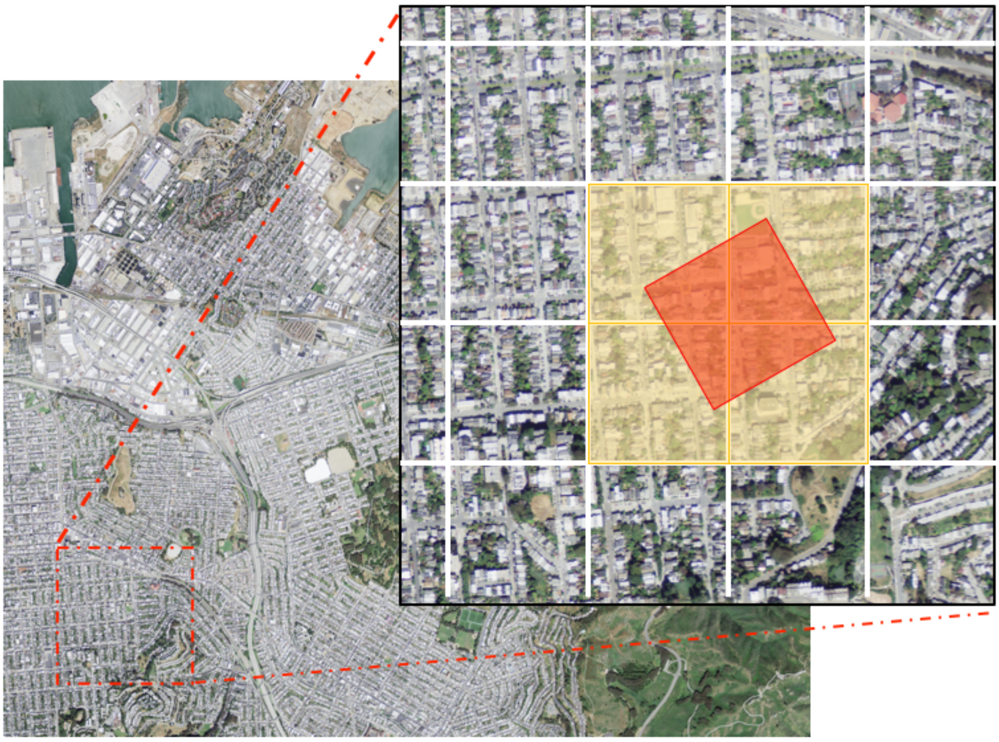
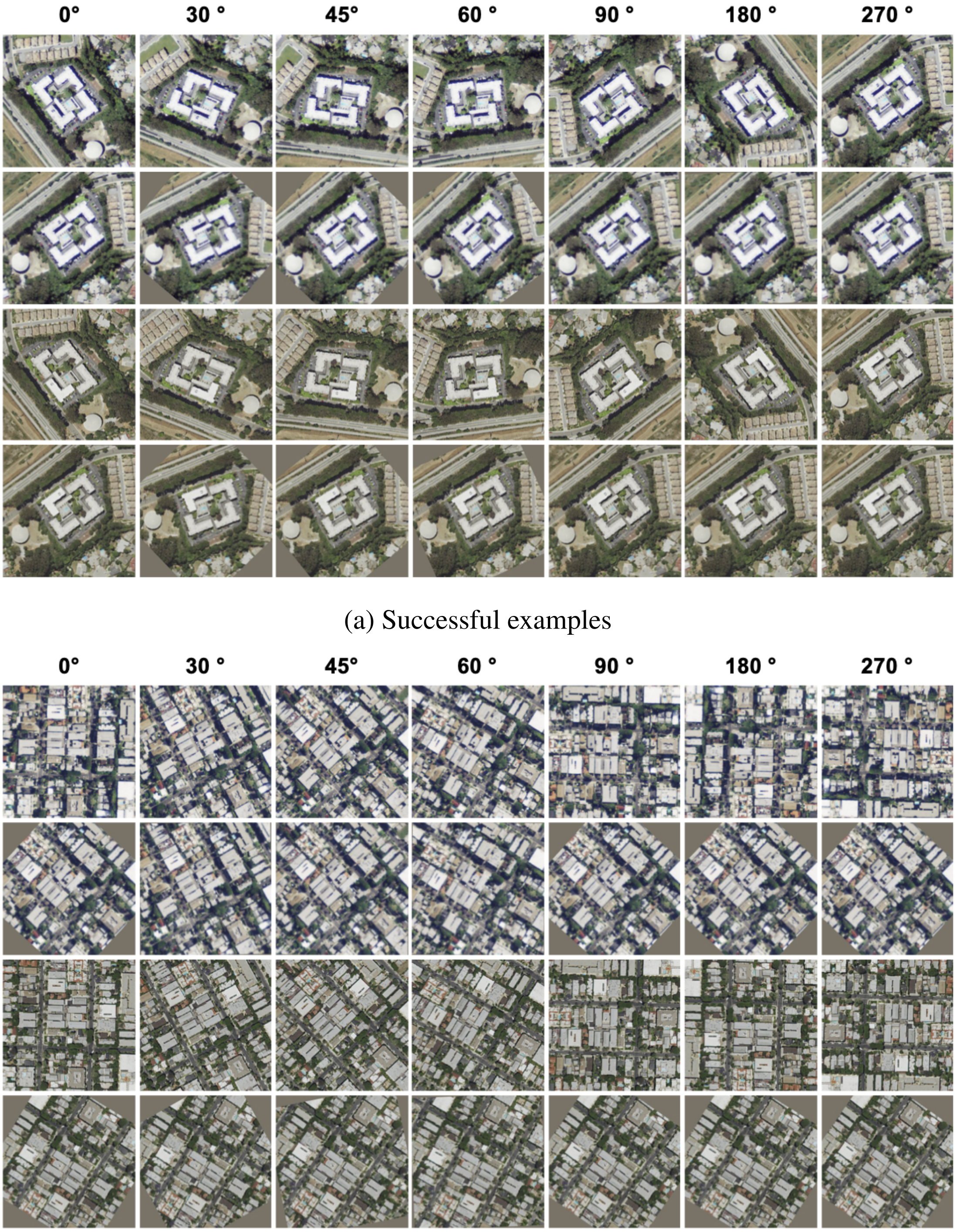


Figure 2.6: A sample query in red and its ground truth in yellow. Geolocalization is successful if the top ranked image in the matched reference set overlaps the query image.

2.5.2 Implementation Details

Siamese networks We fine-tune the Siamese models using an Adam optimizer with a batch size 24. We set the initial learning rate to 10^{-4} for the fc layers and to 10^{-5} for other layers. The learning rate is decayed by 0.1 every 30 epochs. For the RotSiamese network, we set λ_c to 1 and ϕ to either $10°$ or $20°$.

Orientation normalization networks For training the rotation regressor, we use an Adam optimizer with a batch size 24 and an initial learning rate of 2×10^{-5}. We decrease the learning rate by a factor of 10 every 30 epochs. The last Tanh function is scaled by a factor of 1.5π. For ONN training, we set reduce radio rr to $150/224$ and λ_v to 1. For CTONN training, we set rr to $28/224$ and λ_v to 0.1. We experiment with different sets of rotation transformations for training the rotation regressor (parameter K in Section 2.4.3). We consider sets of size 4, 8, and 36 corresponding to multiples



(a) Successful examples

(b) Failed examples

Figure 2.7: (a) Successful and (b) failed orientation normalization examples. The first and third rows contain co-located images from years 2012 and 2014 at various orientations. The second and fourth rows contain the normalized images.

Model	conv3	conv4	conv5	fc6	fc7
Same-time	100	100	91.63	64.00	65.25
Cross-time	76.50	70.63	31.50	19.13	18.50

Table 2.2: Results of performing geolocalization in the SF dataset using features extracted from various layers of a VGG16 network trained on ImageNet. Top: the query and reference images are from the same time; Bottom: they are from different times.

of $90°$, $45°$, and $10°$ respectively. In order to avoid introducing blank regions into the corners of the rotated images, we rotate images of size 370×370 pixels and then extract images of size 256×256 from the center.

Evaluation metrics We consider the geolocalization to be correct if *the top ranked reference tile overlaps the query image*. As shown in Figure 2.6, the ground truth for the red query image is the four reference tiles in yellow since picking any of these tiles would geolocate the query. Using only the top ranked image corresponds to top-1 accuracy which is quite strict. In practice, the top-n ranked images could be marked as candidates and the user could easily make the final selection manually. This would greatly increase performance with modest manual effort. In the case of a correct geolocalization using our method, we assume that image registration could be used to determine the exact location of the query image (such as the geographic coordinates of its corners) using the overlapping reference tiles.

The accuracy for a set of queries is the percentage of successful searches for that set. This is the metric that we report below.

2.5.3 Results

Global vs. local CNN features.

We first compare global versus local features extracted using a VGG16 model trained on the ImageNet dataset. Table 2.2 compares the performance of deep global features extracted from the *fc* layer and deep local features extracted from various *conv* layers. (See Section 2.4.1 for details on how these features are extracted.) These results are for the SF dataset and from when the query and reference tiles have the same orientation.

The top row corresponds to when the query and reference tiles are from the same year and the bottom row to when they are from different years (cross-time). We draw three conclusions. First, as expected, the local features significantly outperform the global features due to the query and ground truth reference tiles only overlapping (see Figure 2.6). Second, the deep local features extracted from *conv3* or *conv4* significantly outperform those from *conv5* especially in the cross-time case. This indicates the features in the final *conv* have possibly become too specialized. (When we train our VGG16 networks on the overhead imagery, the features from *conv4* turn out to be optimal so that is what is used in the experiments below.) Finally, the performance is significantly worse in the cross-time case indicating these features possess limited time-invariance.

Cross-time features and baselines.

Table 2.3 compares our cross-time features trained using the Siamese network to several baselines: NetVLAD [5], fine-tuning image retrieval (FTIR) [90], and SIFT [29]. We also copy the results from the global (VGG_fc) and local (VGG_conv) features extracted using the VGG16 network trained on the ImageNet dataset. The query and reference tiles again have the same orientation. NetVLAD and FTIR are global features and so again perform poorly. Matching using the local SIFT features is also done through voting. While the SIFT features work well in the same-year case, they perform poorly in the cross-year case, indicating they too possess limited time-invariance. Our cross-time deep local features trained using the Siamese network (Section 2.4.2) are shown to outperform all other approaches especially in the cross-year case. The improvement over VGG_conv in the cross-year case in particular demonstrates the effectiveness of our cross-time Siamese training framework.

ONN: orientation invariance.

Table 2.4 shows the results when the query and reference tiles are oriented differently. The columns indicate the difference in orientation between the query and reference: $90°$, $180°$, $270°$, or an arbitrary angle randomly sampled from $(0°, 360°)$. The rows indicate different configurations: no ONN corresponds to no orientation alignment and the other rows indicates the sizes of the sets of rotations the ONN is trained with (K

Method	SF		LA	
	Same year	Cross year	Same year	Cross year
VGG_fc [101]	65.25	18.50	56.89	7.33
NetVLAD [5]	63.63	30.00	65.22	14.00
FTIR [90]	57.11	38.75	55.22	32.33
SIFT [29]	99.75	61.00	100	39.22
VGG_conv [101]	100	70.63	100	47.44
Siamese	100	**82.50**	99.89	**74.11**

Table 2.3: Comparison of our cross-time features (Siamese) with several baselines.

	Test set	SF				LA			
	Rotations	90°	180°	270°	(0°, 360°)	90°	180°	270°	(0°, 360°)
Same time	no ONN	6.50	35.38	6.75	24.25	4.33	27.89	4.11	18.11
	4 classes ONN	99.75	99.63	99.75	69.13	99.56	98.67	99.33	70.89
	8 classes ONN	94.63	94.75	95.13	88.75	97.67	97.78	97.56	96.00
	36 classes ONN	90.38	90.63	90.38	89.25	90.22	91.67	91.67	90.44
Cross time	no ONN	4.75	19.5	3.63	11.63	2.11	14.11	1.22	7.89
	4 classes ONN	79.63	78.13	81.13	35.38	24.78	24.78	26.33	12.56
	8 classes ONN	35.73	36.00	37.38	27.75	14.44	14.33	14.33	12.89
	36 classes ONN	32.5	34.88	32.00	30.13	23.89	26.56	26.11	19.33

Table 2.4: Geolocalization results for when the query and reference have different orientations. See the text for details.

in Section 2.4.3). For example, in the 4 classes cases, the ONN is trained with images at four rotations: 0°, 90°, 180°, and 270°.

We first focus on the same-time case (top of Table 2.4). The no ONN results show just how difficult geolocalization becomes when the query and reference are not oriented the same. Our proposed ONN is shown to significantly improve performance. In particular, the ONN trained with 36 difference rotations achieves around 90% accuracy even in the difficult case of the query having an arbitrary rotation from 0° to 360°. For the fixed rotation cases (90°, 180°, 270°), performance decreases with an increase in training rotation classes. This is because these fixed rotations occur in the training set more often when there are fewer classes.

We now focus on the cross-time case (bottom of Table 2.4). Incorporating the ONN still improves the performance over no ONN but not by as much, especially in the arbitrary orientation case. This demonstrates that the ONN has difficulty normalizing the

Classes	ONN	CTONN	Sia	RotSia(20°)	RotSia(10°)	SF	LA
4	✓		✓			35.38	12.56
	✓			✓		43.5	14.22
		✓		✓		**46.88**	21.89
		✓			✓	40.25	**22.89**
8	✓		✓			27.75	12.89
	✓			✓		31.00	13.33
		✓		✓		**64.88**	37.33
		✓			✓	56.38	**43.33**
36	✓		✓			30.13	19.33
	✓			✓		30.63	20.33
		✓		✓		**62.00**	37.56
		✓			✓	54.75	**45.00**

Table 2.5: Cross-time and arbitrary query orientation results. See the text for details.

orientation of images from different times since that is not what it is trained on.

Figure 2.7 visually illustrates this. The images on the left show the successful normalization of images from different years. The first and third rows show co-located images from different years rotated by varying amounts. These are the inputs to the ONN. The second and fourth rows show the normalized images. These images are similarly oriented both within and between years. In contrast, the images on the right of Figure 2.7 show a failure case. Here, the normalized images on the second and fourth rows are misaligned by $180°$. This is a difficult case, though, even for humans.

CTONN: cross-time orientation invariance.

Table 2.5 shows the results from our CTONN and RotSiamese frameworks when the query and reference are from different times and the query has arbitrary orientation. Remember that the CTONN is trained using differently oriented images from different years and RotSiamese incorporates additional orientation invariance to deal with the noisy CTONN output. (Please refer back to Section 2.4.4 for details.) Sia corresponds to the original Siamese network (Section 2.4.2) and RotSia(20) and RotSia(10) correspond to the RotSiamese networks with $\phi = 20$ and $\phi = 10$ (Section 2.4.4).

The results in Table 2.5 demonstrate that CTONN improves the orientation normalization for images from different times and that the RotSiamese framework learns

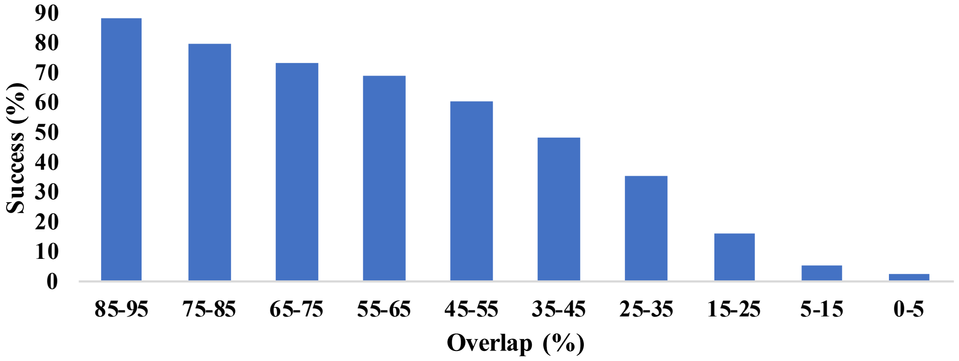


Figure 2.8: Performance versus % overlap between the query and ground truth tiles. See the text for details.

features that are more orientation-invariant. The best results are achieved by combining these two improvements.

Cross-time geolocalization.

Fig. 2.9 shows successful and failed cross time geolocalization examples using our proposed Siamese network framework from section 3.2. These results correspond to table 3. The first row shows the query images which are from 2012 and the second row shows the matched images which are from 2014. The successful examples correspond to when the match overlaps the query. Note how the matched images appear quite different from and often have limited overlap with the queries. This demonstrates the effectiveness of our Siamese network at learning cross time feature representations. In fact, it is even difficult for a human observer to identify the overlap in these successful examples. We challenge the reader to do this. The overlapped images are shown in Fig. 2.15a.

Fig. 2.9b shows examples of failed geolocalizations when the matched (retrieved) image does not overlap the query. These failures are understandable since the images are not distinctive enough to be matched to the co-located images in the reference dataset. They are homogeneous images of water, forest, and dense housing. However, even though geolocalization fails, the matched images are visually and semantically very similar to the queries. This again confirms the effectiveness of our features. It also emphasizes that our problem is different from, and more difficult than, image retrieval.

Cross-time orientation normalization.

In Fig. 2.10, we show the results of our orientation normalization network (ONN) from section 3.3 and our cross time ONN (CTONN) from section 3.4. Again, the goal of these networks it to transform co-located images to the same orientation even if they are from different time. The top row shows a pair of co-located images from 2012 and 2014 rotated by various amounts. The second row shows the results of applying our ONN to the images in the top row and the third row shows the results of our CTONN. We make three observations here. First, our ONN does well at aligning the images at various orientations although there are still slight variations in the results. Second, our ONN mis-aligns this pair of cross time images by approximately 180°. Finally, our CTONN improves upon this and the normalized images are roughly in the same orientation. This demonstrates the effectiveness of our CTONN at aligning images from different time to the same orientation. The geolocalization performance of our ONN and CTONN is compared in table 5.

Cross-time and orientation-invariant image geolocalization.

To understand how our geolocalization framework performs on different types of terrain, we visualize the performance of our different approaches in the SF area in Fig. 2.11 and the LA area in Fig. 2.12. Each circle in these figures represents the location of query image from 2012 with arbitrary orientation. (Our queries are square regions. The circles correspond to the centers of these regions.) A green circle indicates a successful geolocalization in which the matched image from the 2014 reference overlaps the query image. A red circle indicates a failure. Similar to the results in Fig. 2.9 above, we expect our framework to fail for homogeneous, non-distinctive regions.

Results from two scenes from each of SF and LA are shown. The images on the left show the results when the query is oriented the same as the reference imagery. We consider this the easy case and apply the Siamese network from section 3.2. The results on the right show the results when the query is arbitrarily oriented. We consider this the difficult case. These results are produced using our combined CTONN and RotSiamese frameworks (section 3.4).

As we can see from the results in Fig. 2.11 and Fig. 2.12, our proposed framework achieves good performance when the query is at the same orientation as the reference imagery. The few failure cases here are typically when the query terrain is non-distinctive such as over water. Looking at the second image from SF, we notice that our framework can geolocalize query images of forest regions when the queries are oriented in the same direction as the reference but that this performance deteriorates when the query regions are arbitrarily oriented. This shows it is difficult to normalize the orientation of images of forest which makes sense because there is little directional information in these images. For the LA area shown in Fig. 2.12, failure at locating forest images can also be observed. Our framework is able to successfully estimate the location of dense buildings and streets in LA.

Fig. 2.13 and Fig. 2.14 show successful and failed examples of cross time and arbitrary orientation geolocalization examples from the SF and LA areas respectively. The significant visual differences caused by cross time domain and orientation can be observed which makes geolocalization more challenging. The shape of the buildings and the streets are obscured by trees and different viewpoints, especially for LA. However, our method is able to provide a good solution. Even in the failure cases, our framework returns visually similar images to the reference set even if they are not co-located with the queries. This again demonstrates the effectiveness of our methods. We again propose a challenge to the reader to find the overlapping regions for the successful examples. The overlapping regions for these examples are shown in Fig. 2.15b and Fig. 2.15c.

2.6 Discussion

Limitations:

We note that our orientation normalization framework only works if the normalized images can be matched. It cannot improve over the case where the query and reference have the same orientation.

Our performance will of course depend on the content of the query. If the query is not distinctive, such as a homogeneous image of water, forest, or even dense housing, our framework will likely fail due to there being tiles in the reference which, particularly in the cross-year case, are more similar to the query than the ground truth. But, any image-based approach would fail in this case. We note, though, that even when we fail to geolocate the query images, the top matches are visually and semantically very similar. This again emphasizes that performing effective similarity-based image retrieval is not sufficient for our problem.

Our approach currently uses co-located image pairs from the query and reference datasets when training the cross-time and orientation-invariant components. Such pairs will not always be available and so this is another limitation.

Finally, our framework will fail when we do not have reference imagery for the query location. But, high-resolution overhead imagery is available for most if not all of the Earth.

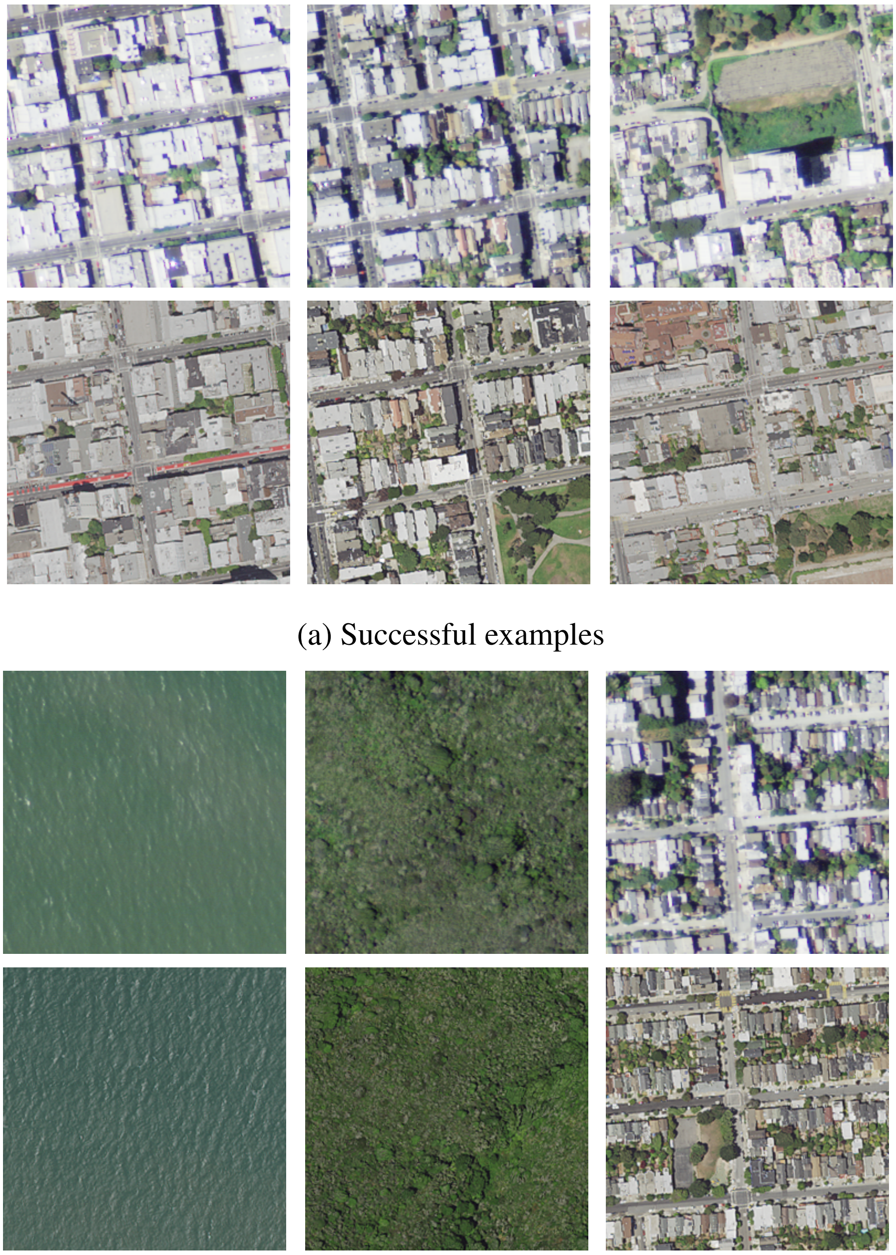
Scalability to partial overlap

Finally, we explore how sensitive our approach is to the overlap between the query and the reference tiles. Figure 2.8 shows the performance as a function of % overlap. *Success here means that a ground truth tile that overlaps the query by a certain amount is in the top n matches where n is the number of ground truth tiles.* (The ground truth tiles are those that overlap the query.) Note, though, that our geolocalization framework only requires that *one* of the ground truth tiles is the top match, not that all of the ground truth tiles are in the top matches. Since we assume a set of contiguous references tiles, as the overlap between the query and any one ground truth tile decreases, the overlap with another ground truth tile necessarily increases (see Figure 2.6). At least one reference tile overlaps with the query image more than 25%.

Figure 2.8 shows that, as expected, the ability of our matching framework to retrieve a ground truth tile decreases as the overlap decreases. The features are only so local due to the spatial entanglement of the convolutional maps.

2.7 Conclusion

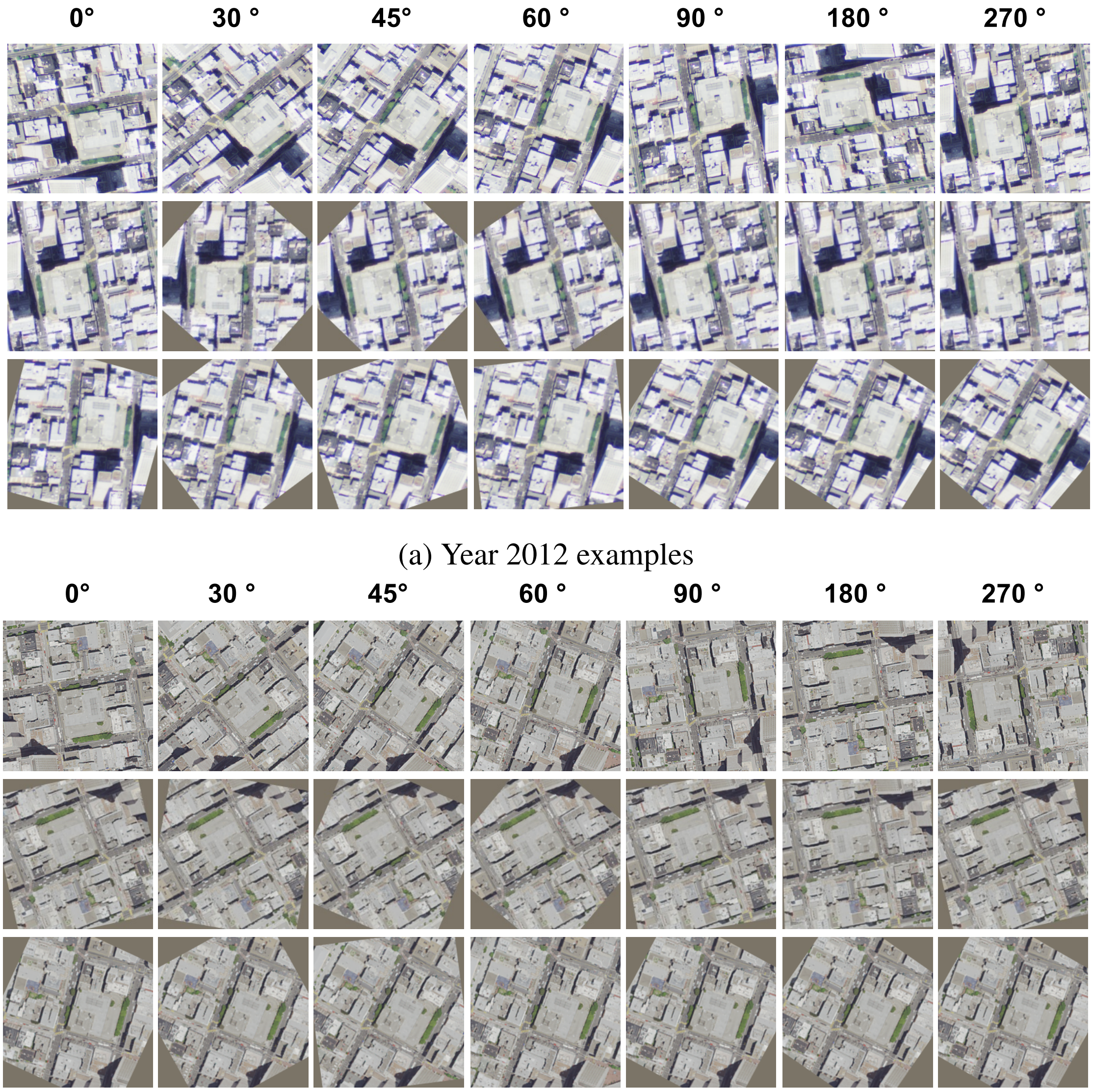
We perform large-scale overhead image geolocalization by matching a query image to wide-area reference imagery with known location. We demonstrate that local features, particularly those extracted using CNNs, are more effective than global features due to the partial overlap of the query and reference tiles. We develop several technical innovations to deal with the real but challenging cases of when the query and reference are from different times and when the query has an arbitrary orientation. We demonstrate the effectiveness of these innovations on two large datasets of high-resolution aerial imagery.



(a) Successful examples

(b) Failed examples

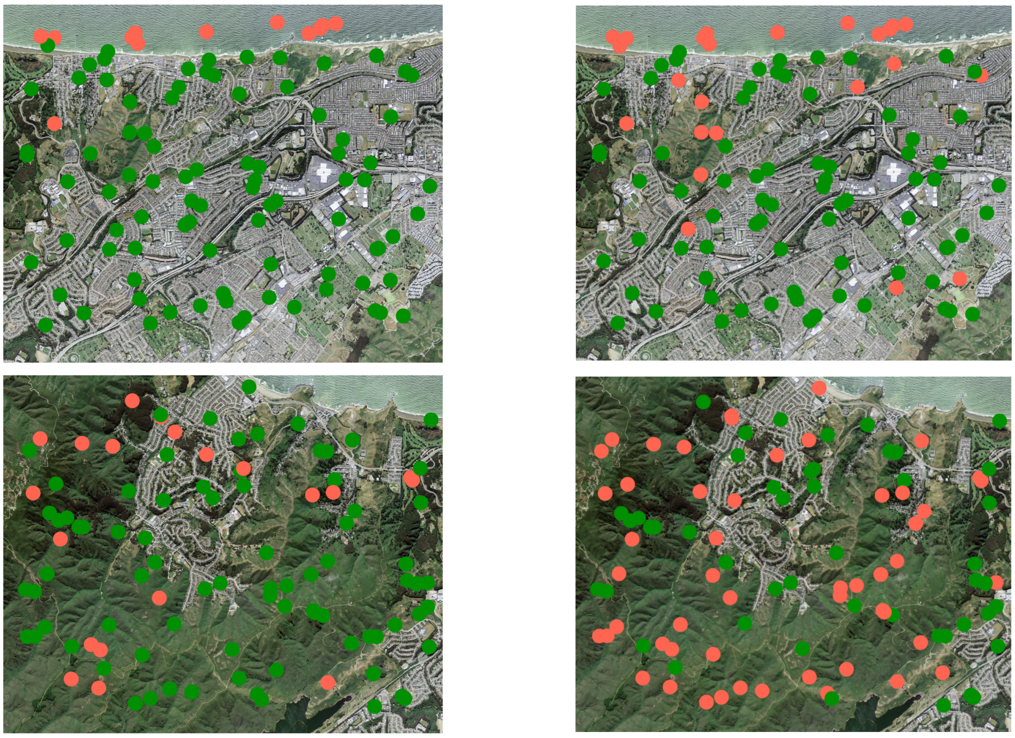
Figure 2.9: (a) Successful and (b) failed cross-time geolocalization examples from the SF area. The first row contains query images from 2012 and the second row contains the matched images from 2014. We challenge the reader to determine the overlap in the successful examples. The overlapped images are shown in Fig. 2.15a.



(a) Year 2012 examples

(b) Year 2014 examples

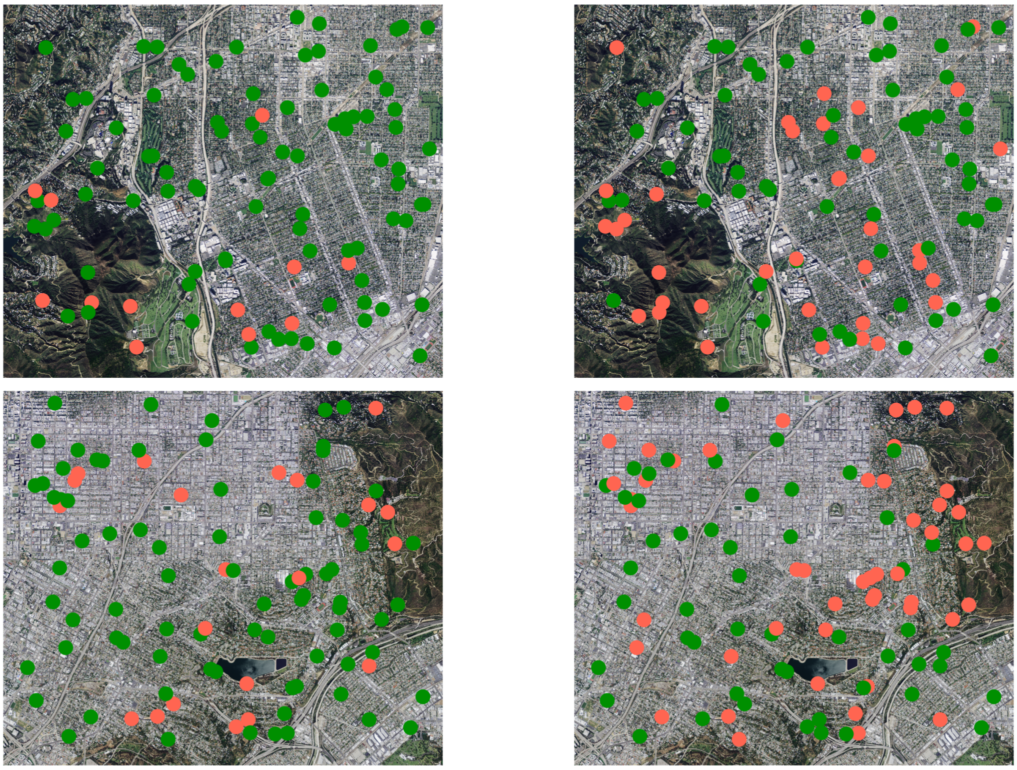
Figure 2.10: (a) Year 2012 and (b) Year 2014 orientation normalization examples. The first row contains a co-located pair of images from 2012 and 2014 at various orientations. The second row contains the images normalized using our orientation normalization network (ONN). The third row contains the images normalized using our cross-time ONN (CTONN).



(a) Query without rotation

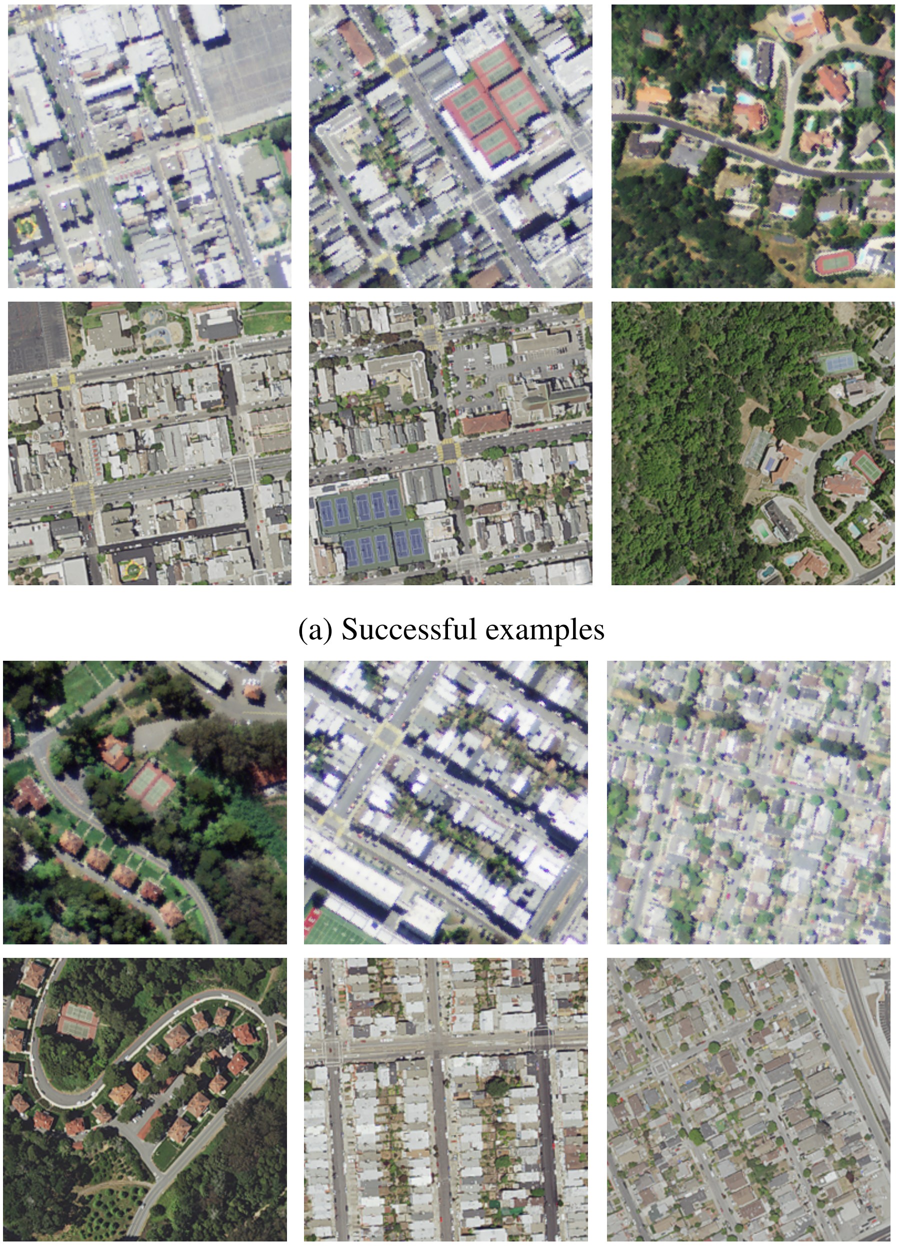
(b) Query with arbitrary orientation

Figure 2.11: Examples of cross-time geolocalization for two NAIP scenes from SF area. Each circle corresponds to a query image with successful (green) or failed (red) geolocalization result. (a) Query images without rotation where the Siamese framework is used. (b) Query images with arbitrary orientation where the combined CTONN and RotSiamese frameworks are used.



(a) Query without rotation (b) Query with arbitrary orientation

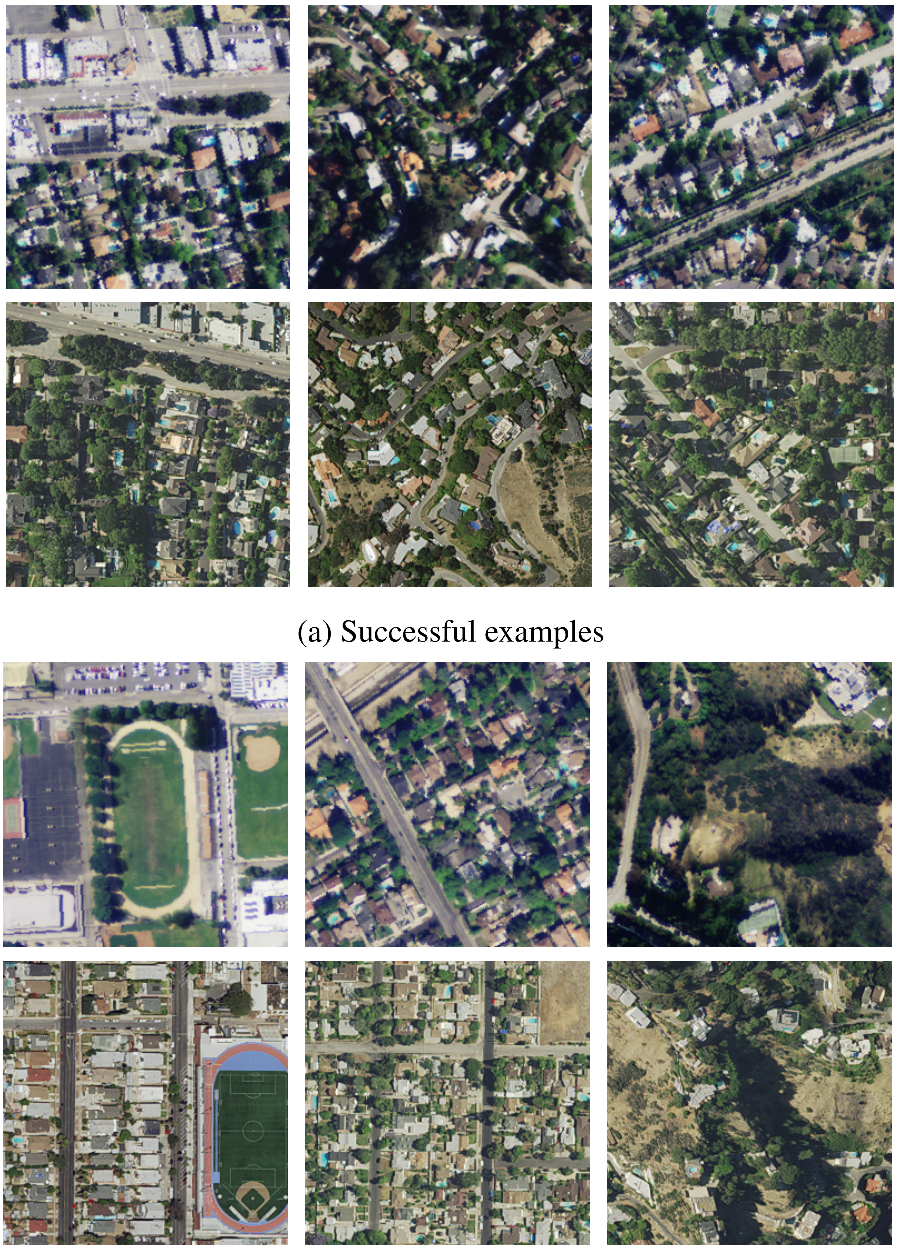
Figure 2.12: Examples of cross-time geolocalization for two NAIP scenes from LA area. Each circle corresponds to a query image with successful (green) or failed (red) geolocalization result. (a) Query images without rotation where the Siamese framework is used. (b) Query images with arbitrary orientation where the combined CTONN and RotSiamese frameworks are used.



(a) Successful examples

(b) Failed examples

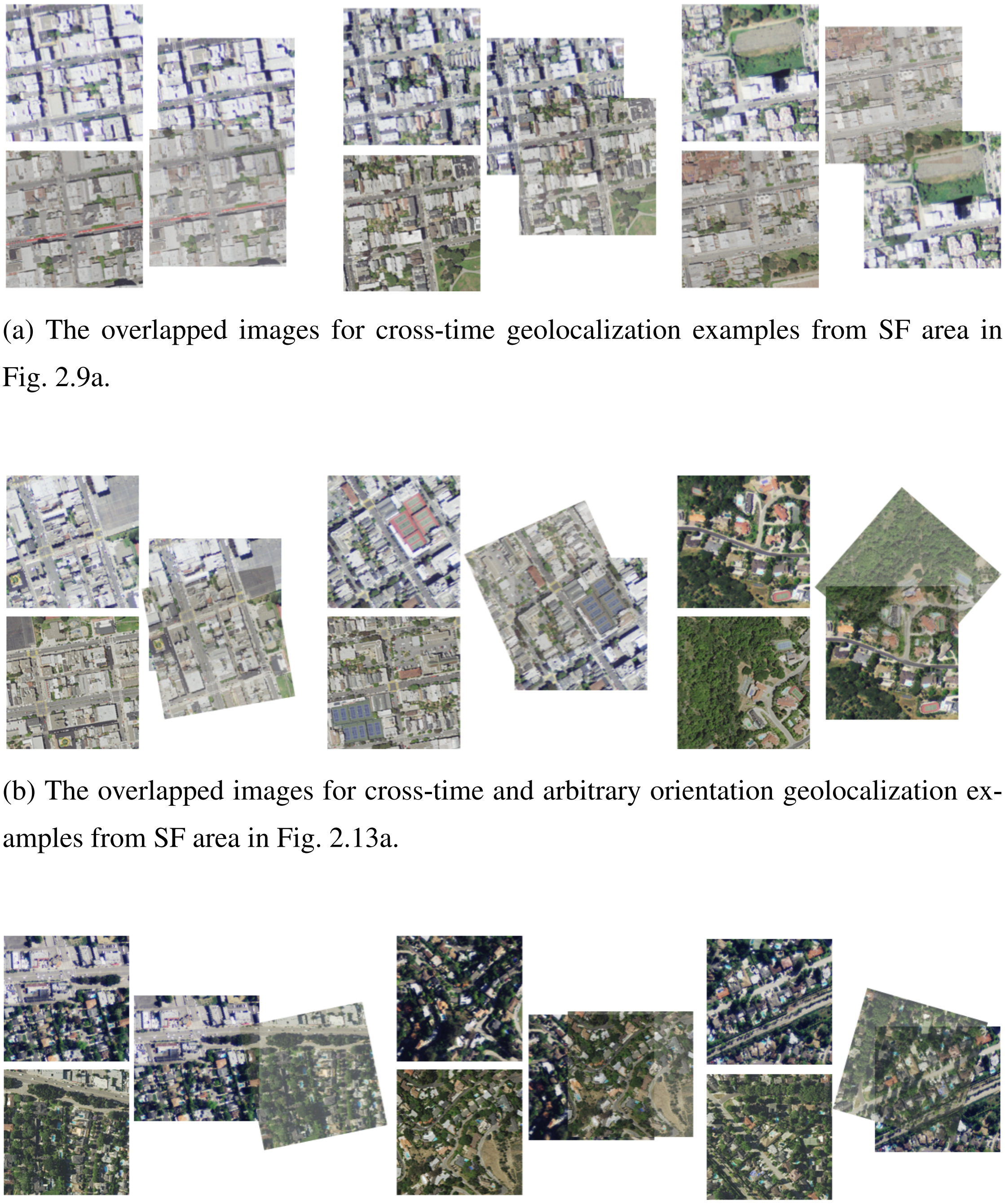
Figure 2.13: Examples of (a) Successful and (b) failed geolocalization for cross-time and arbitrary orientation from SF area. Each column consists of a pair of images with query image (top) with arbitrary angle from 2012 and the matched image (bottom) from 2014 after applying the CTONN and RotSiamese frameworks to the query images. Examples of the overlapping regions are shown in Fig. 2.15b.



(a) Successful examples

(b) Failed examples

Figure 2.14: Examples of (a) Successful and (b) failed geolocalization for cross-time and arbitrary orientation from LA area. Each column consists of a pair of images with query image (top) with arbitrary angle from 2012 and the matched image (bottom) from 2014 after applying the CTONN and RotSiamese frameworks to the query images. Examples of the overlapping regions are shown in Fig. 2.15c.



(a) The overlapped images for cross-time geolocalization examples from SF area in Fig. 2.9a.

(b) The overlapped images for cross-time and arbitrary orientation geolocalization examples from SF area in Fig. 2.13a.

(c) The overlapped images for cross-time and arbitrary orientation geolocalization examples from LA area in Fig. 2.14a.

Figure 2.15: Examples of overlapping regions for query and matched images of successful geolocalization results in (a) Fig. 2.9, (b) Fig. 2.13 and (c) Fig. 2.14.

Chapter 3

Image Search with Text Feedback by Additive Attention Compositional Learning

3.1 Overview

Effective fashion image retrieval with text feedback stands to impact a range of real-world applications, such as e-commerce. Given a source image and text feedback that describes the desired modifications to that image, the goal is to retrieve the target images that resemble the source yet satisfy the given modifications by composing a multi-modal (image-text) query. We propose a novel solution to this problem, Additive Attention Compositional Learning (AACL), that uses a multi-modal transformer-based architecture and effectively models the image-text contexts. Specifically, we propose a novel image-text composition module based on additive attention that can be seamlessly plugged into deep neural networks. We also introduce a new challenging benchmark derived from the Shopping100k dataset. AACL is evaluated on three large-scale datasets (FashionIQ, Fashion200k, and Shopping100k), each with strong baselines. Extensive experiments show that AACL achieves new state-of-the-art results on all three datasets.

3.2 Introduction

Image retrieval is a fundamental task in computer vision and serves as the cornerstone for a wide range of applications such as fashion retrieval [70, 98], geolocalization [68, 105], and face recognition [103]. There are several ways to formulate the search query such as keywords [1, 138], a query image [130, 117], or even a sketch [36, 57, 133, 13, 14, 93]. However, a core challenge in traditional image retrieval is that it is difficult for the user to refine the retrieved items based on their intentions. A range of approaches to incorporate user feedback to refine the retrieved images have been explored. Combining natural language feedback with a query image is a particularly promising framework since it provides a natural and flexible way for users to convey the image modifications that they have in mind.

In this work, we investigate image retrieval with text feedback where the goal is to retrieve images that are similar to a query image but incorporate the modifications described by the text. Such multi-modal and complementary input provides users with a powerful and intuitive visual search experience. However, as a multi-modal learning

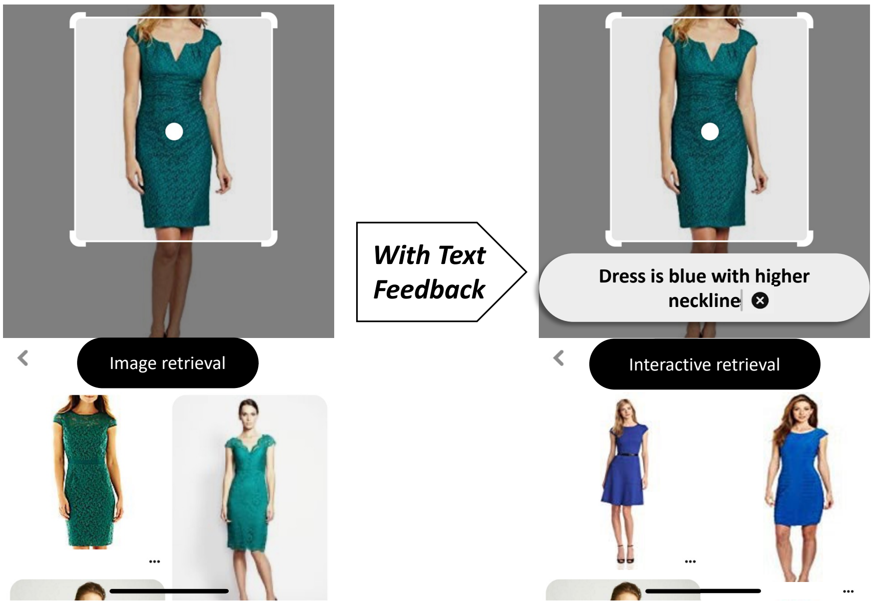


Figure 3.1: We consider the task of retrieving new images that resemble the reference image while changing certain aspects as specified by text. Best viewed in color.

problem, it requires the synergistic understanding of both visual and linguistic content which can be a challenge. While image search with text feedback lies at the intersection of vision and language analysis, it differs from other extensively studied vision-and-language tasks, such as image-text matching [65, 63, 140, 50], image captioning [91, 84, 25], and visual question answering [38, 52, 18, 16]. This difference stems from the significant challenge of learning a *composite representation* that jointly captures the *visual* content of the query image and the *linguistic* information in the accompanying text to match the target image of interest.

A fundamental challenge in image-text compositonal learning is characterizing global concepts from the query image and text representation simultaneously. For instance, when the text describes a modification to the color and neckline of a dress in a query image, the composition module should capture the concept of transforming the color and neckline, but it should also preserve the other visual concepts such as the trim, and material of the dress (Figure 3.1). Another challenge is how to *selectively* modify the query image representation using the captured contextual information so that it is close to the target image representation in the latent space.

We propose a novel transformer-based Additive Attention Compositional Learning

(AACL) model to address these challenges. The key idea is that we learn a contextual vector from the joint visiolinguistic representation. AACL then selectively modifies the query image tokens using the global context vector such that the composite features preserve the visual content of the image that should not be changed while transforming the relevant content according to the accompanying text.

We empirically compare our AACL approach with the state-of-the-art (SOTA) methods for visual search with text feedback on three large-scale fashion datasets: FashionIQ [39], Fashion200k [41], and a new challenging benchmark derived from Shopping100k [2]. We show that our proposed compositional learning method outperforms existing methods on all three datasets.

We make the following fundamental contributions:

- We propose a novel multi-modal additive attention layer capable of learning a global context vector which is used to selectively modify the image representation in an efficient way.

- We develop a fully transformer-based model for the challenging task of visual search with text feedback and demonstrate that it achieves state-of-the-art performance through extensive experiments on several large-scale fashion datasets.

- We create a new image-text retrieval dataset derived from Shopping100k. This new dataset features a wider range of fashion categories and attributes, resulting in an additional challenging benchmark for the research community.

3.3 Related Work

3.3.1 Image Retrieval with Text Feedback

Image retrieval with text feedback has been of interest to the computer vision research community for some time and a number of efforts (e.g., [4, 76, 110]) have investigated effective ways to combine image and text representations. The text feedback can be provided in various ways, including absolute attributes (e.g., “red”) [1, 138, 41], simple relative attributes (e.g., “more red”) [85, 60, 131], or full natural language phrases [110, 3, 51, 20, 30, 100, 55]. Natural language is the preferred method of interaction between humans and computers in contemporary search engines. For image search in particular, it allows a user to convey detailed and precise specifications or modifications in a very natural way. We therefore focus on query-based image search with accompanying natural language phrases.

Previous methods [3, 19, 55, 30, 100] for image retrieval with text feedback rely heavily on convolution to aggregate features. In contrast, ours is the *first approach to efficiently learn features globally via attention.* Previous works have also relied on complicated hierarchical feature aggregation [20, 51], multiple forms of text feedback [20, 3], or multiple loss functions [20, 51, 3]. The winning solutions [55, 56, 99] for the FashionIQ 2020 challenge—an interactive image retrieval challenge—employed common performance boosting techniques such as careful hyperparameter tuning and model ensembles to improve the results. In contrast, AACL focuses on the *design of the image-text composition module* and achieves state-of-the-art performance via feature fusion in one step, which is more efficient and easier to adapt to other frameworks.

3.3.2 Image-Text Composition

While there has been much effort and different kinds of methods proposed to achieve the top scores on benchmarks involving image and text, relatively few have focused on the image-text composition module itself. In [54], the authors propose a multi-modal residual network (MRN) that learns representations by fusing visual and textual features through element-wise multiplication and residual learning. FiLM [87] utilizes a linear modulation component in which text information modifies the image representation via a feature-wise affine transformation. Vo *et al.* proposed TIRG [110], which uses a gating mechanism to determine the channels of the image representation that should be modified by the conditioning text. In ComposeAE [3], a complex embedding space that semantically ties the representations from text and image modalities is designed. Recently, MAAF [30] improved multi-modal image search via a Modality-Agnostic Attention Fusion model. This model uses a dot product attention mechanism as found in the standard transformer architecture. Additionally, resolution-wise pooling is proposed to aggregate fine-grained features from a ResNet [43] CNN. RTIC [100] consists of a residual text and image composer to encode the errors between the source and target images in the latent space and includes a graph convolutional network for regularization. Our work differs from these composition modules in that we utilize a novel image and text composition module via additive attention [6, 79] to model global contexts. Furthermore, we use an element-wise product to model the interaction between the global context and each input token, which both greatly reduces the computational cost and effectively captures the contextual information [54, 55, 124].

3.3.3 Attention Mechanism

The concept of attention has gained popularity recently in neural networks as it allows the models to learn representations from different modalities [54, 48, 30, 20, 4]. The two most commonly used attention functions are additive [6], and dot-product (multiplicative) attention [109]. Dot-product attention has a drawback, however, in that it has to attend to all the tokens on the source side for each target token, which is expensive and can potentially be impractical for longer sequences. Additive attention has been shown experimentally to achieve higher accuracy than multiplicative attention in some scenarios [79, 124]. Inspired by this, we propose an *additive attention composition module* for feature fusion.

3.4 Method

Figure 3.2 presents the overall architecture of our Additive Attention Compositional Learning (AACL) framework. Given a source image x and text feedback t as the input query, the goal of AACL is to learn a composite representation o_{xt} that can be used to retrieve relevant images y from a target database. AACL contains three key components: (1) an image encoder for visual semantic representation learning, (2) a text encoder for natural language representation learning, and (3) an additive attention composition module that modifies the source image representation according to the text representation. In contrast to other approaches that use multiple stages of feature composition and matching (e.g., [20]), AACL does this in one stage using the final output of the image and text encoders.

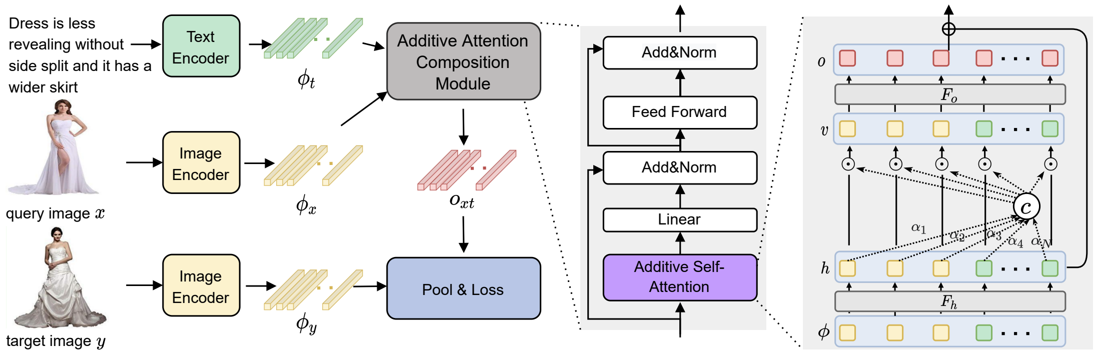


Figure 3.2: Overview of our Additive Attention Compositional Learning framework. Given a pair of query image and text as input, our goal is to learn a composite representation that aligns to the target image representation. AACL contains three major components: an image encoder (Sec. 3.4.1), a text encoder (Sec. 3.4.1), and an Additive Attention Composition Module (Sec 3.4.2) that can be plugged into different models for feature fusion. "$\odot$" represents Hadamard product.

In the following, we first provide an overview of the two encoders in Section 3.4.1. We then detail our novel composition module in Section 3.4.2 and our model optimization in Section 3.4.3.

3.4.1 Image and text representation

Image Representation: We employ a Swin Transformer [74] to derive a discriminative representation of the visual content of an image. As a transformer inherently learns visual concepts of increasing abstraction in a compositional, hierarchical order, we conjecture that image features from the final layer may not fully capture the visual information of the lower levels. We thus concatenate image tokens extracted from the final (Stage 4) and penultimate (Stage 3) layers of the Swin Transformer. Unless otherwise specified, our model uses these $49 + 49 = 98$ image tokens for multi-level image understanding. A learned linear projection maps each image token to d dimensions so that the final image representation is $\phi_x \in \mathbb{R}^{98 \times d}$.

Text Representation: The DistilBERT language representation model [94] is used to encode the semantics of the accompanying text. DistilBERT naturally yields m tokens for the input words, namely the hidden states of the last layer of the model. We concatenate these tokens to form the final text representation $\phi_t \in \mathbb{R}^{m \times d}$.

3.4.2 Additive Attention Composition Module

In order to jointly represent the image and text components of the query, we seek to transform the visual features conditioned on language semantics. To accomplish this, we propose an *additive attention composition module* for feature fusion. This module consists of multiple composition blocks that each employ additive self-attention to learn a context vector which then selectively modifies the joint visiolinguistic representation. The final output of these blocks yields a modified image representation that is meant to faithfully capture the input image and text information.

Visiolinguistic Representation: In order to obtain the input representation for our first composition block, the image tokens ϕ_x and text tokens ϕ_t are concatenated to obtain the visiolinguistic representation $\phi = [\phi_x, \phi_t]$. The final representation is denoted as $\phi \in \mathbb{R}^{N \times d}$, where N is the combined count of image and text tokens.

Composition Block: Following the standard transformer architecture [109], the additive attention composition module is composed of a stack of L identical blocks with multiple heads. Different attention heads use the same formulation but different parameters, which allows the model to jointly attend to information from different representation subspaces at different positions. Each block has an additive self-attention layer followed by a linear layer and a feed-forward neural network. We also employ a residual connection and layer normalization after these linear and feed-forward components.

Additive Self-Attention Layer: In order to discover the latent relationships essential for learning the transformation, we use the additive attention mechanism to learn a context vector c, then selectively suppress and highlight the representations from each token. Similar to [124], we first use a linear transformation layer to transform the input sequence into the hidden states: $h = \mathcal{F}_h(\phi_i), i \in N$. The context vector c that is learned to modify each token is generated as a weighted sum of these tokens h_i:

$$c = \sum_{i=1}^{N} \alpha_i h_i, \tag{3.1}$$

The weight α_i of each token h_i is computed by

$$\alpha_i = \frac{\exp\left(\mathbf{w}_h^T \mathbf{h}_i / \sqrt{d}\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{w}_h^T \mathbf{h}_j / \sqrt{d}\right)}. \tag{3.2}$$

where $\mathbf{w}_h \in \mathbb{R}^d$ is learned during the training process, and $\mathbf{w}_h^T \mathbf{h}_j$ scores how much each input token contributes to the global context.

Next, to selectively suppress and highlight the visual content in h, a Hadamard product is introduced to reuse the global contextual information, which is motivated by its effectiveness in modeling the nonlinear relationship between two vectors [113, 124, 45]. It is formulated as $v_i = c \odot h_i$. Another linear transformation layer $\mathcal{F}_o$ is applied to each token v_i to learn its hidden representation. To form the final output of the additive attention layer, we add the hidden states h_i that capture relevant source-side information to the transformed latent features. The final output of the additive self attention layer is:

$$o_i = h_i + \mathcal{F}_o\left(c \odot h_i\right) \tag{3.3}$$

3.4.3 Deep Metric Learning

Our objective during training is to push the “modified” image representation ϕ_{xt} and the target image representations ϕ_y closer, while pulling apart the representations of dissimilar images. A batch-based classification loss as in [110, 3, 30] is used to train the model as early experiments showed that the triplet loss performs worse for the Recall@k metric. Each batch is constructed from N pairs of a query (image and text) and its corresponding target image.

$$L = \frac{1}{B} \sum_{i=1}^{B} - \log \left\{ \frac{\exp\left\{\kappa\left(\phi_y, \phi_{xt}\right)\right\}}{\sum_{j=1}^{B} \exp\left\{\kappa\left(\phi_y, \phi_{xt}\right)\right\}} \right\} \tag{3.4}$$

where B is the batch size and κ is a similarity kernel that is implemented as the dot product in our experiments.

3.5 Experiments

3.5.1 Experimental Setup

Datasets: We evaluate our model on three datasets—FashionIQ, Fashion200k and our modified version of Shopping100k—in order to validate its ability to generalize to a variety of natural language expressions. We provide details of these datasets in Sections 3.5.2, 3.5.3, and 3.5.4, respectively.

Implementation Details: We use the PyTorch deep learning framework to conduct all our experiments. The Swin Transformer [74] is used as the backbone for the image encoder. The transformer model is initialized using weights first pre-trained on ImageNet-22K and then fine-tuned on ImageNet-1K [27].

We extract sequences of 1024-dimensional tokens from Stages 3 and 4 of the model and then project the tokens to d dimensions, which for our experiments is 768. We learn the text embedding using a pre-trained DistilBERT model [94], which yields a 768-dimensional token for each input word. The original BERT model is pre-trained on BooksCorpus (800M words) and English Wikipedia (2,500M words) [28]. We employ 3 additive attention composition blocks and 8 parallel attention heads for each block. For training, we use SGD optimization with a learning rate of 0.035. We train all models using 4 GPUs with a batch size of 32 per GPU. For FashionIQ, we employ a learning rate decay of 0.1 every 10 epochs for 60 epochs. For Fashion200k and our modified Shopping100k, we use the same decay value but every 30 epochs with a total of 100 epochs. We report the average and standard deviation of five trials for all our experiments to obtain more meaningful results.

Evaluation Metric: For evaluation we adopt Recall@K (denoted as R@K for short), a standard metric in retrieval.

Compared Methods: We compare the results of AACL with several methods, namely: FiLM, MRN, TIRG, ComposeAE, MAAF and RTIC. We explained them briefly in Section 3.3.2.

Model	Shirt		Dress		Toptee		Average
	R@10	R@50	R@10	R@50	R@10	R@50	(R@10 + R@50)/2
MRN [54]	15.88	34.33	12.32	32.18	18.11	36.33	24.86
FiLM [87]	15.04	34.09	14.23	33.34	17.30	37.68	25.28
TIRG [110]	16.12	37.69	19.15	43.01	21.21	47.08	30.71
ComposeAE [3]	9.96	25.14	10.77	28.29	12.74	30.79	19.61
MAAF [30]	21.30	44.20	23.80	48.60	27.90	53.60	36.57
RTIC [100]	22.03	45.29	27.37	52.95	27.33	53.60	38.10
TIRG*	21.38±0.54	46.28±0.78	25.82±0.39	53.21±0.33	26.73±0.72	53.17±0.29	37.77±0.21
MAAF*	23.55±0.31	46.38±1.34	28.75±0.63	54.48±0.49	29.70±0.45	55.84±0.87	39.78±0.68
RTIC*	23.03±0.63	46.68±0.52	26.86±0.74	52.80±0.61	27.21±0.89	53.24±0.66	38.31±0.67
AACL	**24.82±0.62**	**48.85±0.77**	**29.89±0.65**	**55.85±0.87**	**30.88±1.2**	**56.85±1.16**	**41.19±0.88**

Table 3.1: Comparison of image search with text feedback on FashionIQ. Averaged R@10/50 computed over all three categories. * denotes results obtained with the same image encoder and text encoder as AACL.

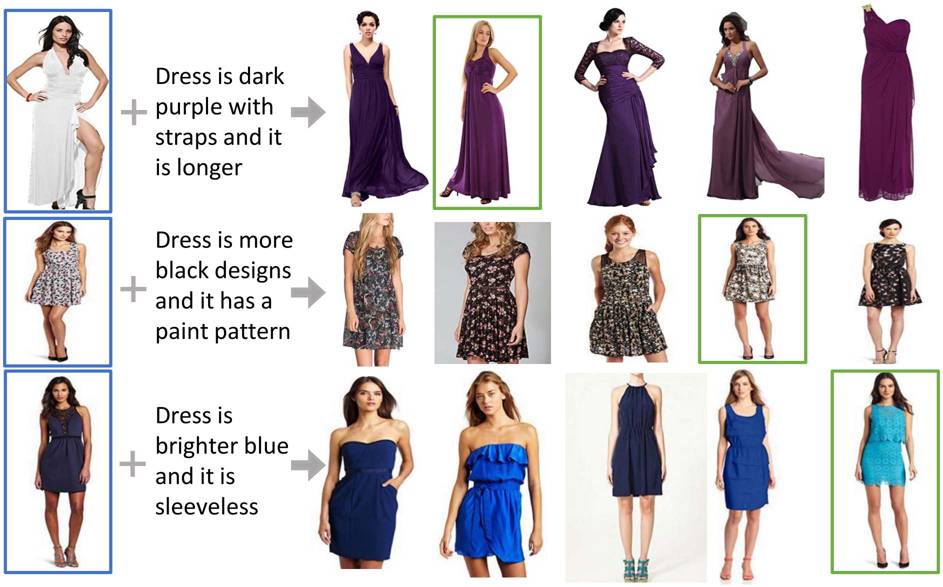

Dress is dark purple with straps and it is longer
Dress is more black designs and it has a paint pattern
Dress is brighter blue and it is sleeveless

Figure 3.3: Qualitative results of AACL on FashionIQ dataset. Blue and green box indicate query and target images, respectively.

3.5.2 FashionIQ

FashionIQ [39] is a natural language based interactive fashion product retrieval dataset. It contains 77,684 images crawled from Amazon.com, covering three categories: Dresses, Tops&Tees and Shirts. Among the 46,609 training images, there are 18,000 image pairs. Each pair is accompanied by on average two natural language sentences that describe one or multiple visual properties to modify in the reference image,

Model	R@1	R@10	R@50
FiLM [87]	12.9	39.5	61.9
MRN [54]	13.4	40.0	61.9
TIRG [110]	14.1	42.5	63.8
ComposeAE [3]	16.5	45.4	63.1
DCNet [55]	–	46.9	67.6
MAAF [30]	18.94	–	–
TIRG*	17.22±0.39	56.52±1.85	75.60±0.09
MAAF*	17.79±0.98	57.57±0.98	77.51±0.63
RTIC*	17.05±0.96	54.65±0.79	75.54±1.63
AACL	**19.64±1.66**	**58.85±1.01**	**78.86±0.43**

Table 3.2: Comparison of image search with text feedback on Fashion200k dataset. * denotes our implementation results obtained with the same image encoder and text encoder as AACL.

such as *"is shiny"* or *"is blue in color and floral, and with white base"*. We follow the same evaluation protocol as [39], using the same training split and evaluating on the validation set. We report results on individual categories, as well as the average results over all three.

Table 3.1 compares the performance of AACL and the other methods on FashionIQ. We observe that AACL is superior to all reported results by a large margin (top half). AACL even outperforms methods that include factors other than the composition module itself, such as the target image captions, model ensembles, and additional joint loss functions [3]. We further note that AACL is actually complementary to some of these methods and could, in fact, be used as their composition modules. For a like-to-like fair comparison, we also reproduced the best competitors, focusing on just the composition module itself. That is, we utilized the same image and text encoders—namely, Swin Transformer and DistilBERT—and the same optimizer. In this scenario AACL surpasses TIRG, RTIC, and MAAF by an overall margin of 3.42%, 2.88% and 1.41% respectively in average R@10 and R@50 scores. Figure 3.3 presents our qualitative results on FashionIQ. These results demonstrate that our model can handle complex and realistic text descriptions. We also observe that our model can jointly comprehend global appearance (e.g., colors, material), as well as local fine-grained details (e.g., straps and neckline, length of sleeves), for image search.

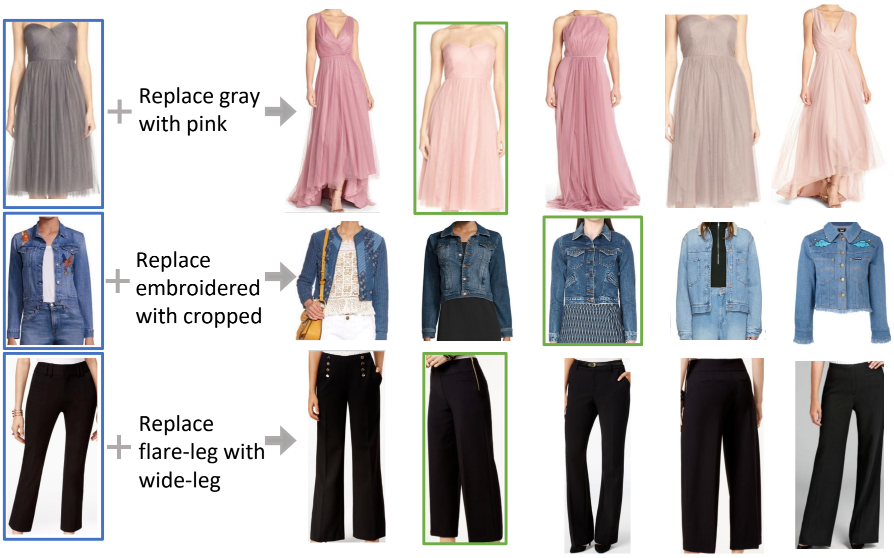

Replace gray with pink
Replace embroidered with cropped
Replace flare-leg with wide-leg

Figure 3.4: Qualitative results of AACL on Fashion200k dataset. Blue and green box indicate query and target images, respectively.

Jacket	Shirt	T-shirt	Jumper	Shorts	Trouser	Jean	Swim	Bottoms[1]	Skirt	Dress
7,528	14,853	22,071	11,797	5,099	4,630	6,229	5,497	3,726	2,528	12,119

Table 3.3: Number of images in select categories (count > 2k) in Shopping100k dataset.

3.5.3 Fashion200k

Fashion200k [41] is a large-scale fashion dataset crawled from multiple online shopping websites. It contains more than 200k fashion images collected for attribute-based product retrieval. It also covers a diverse range of fashion concepts, with a total vocabulary size of 5,590. Each image is tagged with descriptive text corresponding to a product description, such as *"beige v-neck bell-sleeve top"*. Following [110], we use the training split of 172,049 images for training and the test set of 33,480 test queries for evaluation. During training, pairwise images with attribute-like modification texts are generated by comparing their product descriptions on-the-fly, e.g., *"replace black with blue"* or *"replace mini with midi"*.

Table 3.2 shows our model achieves compelling results compared to other methods, most notably for R@1 where AACL outperforms the best competitor MAAF by a rel-

[1] Full name of category "Bottoms" is "Tracksuit Bottoms".

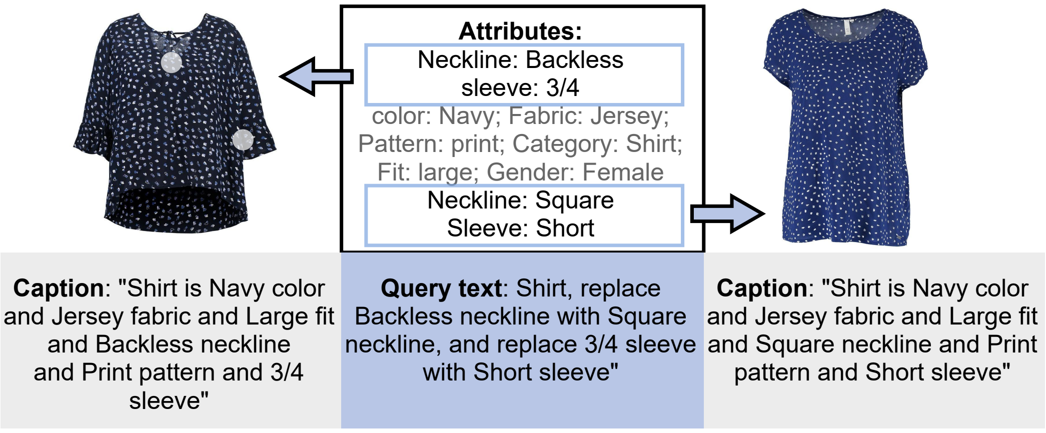

Attributes:
Neckline: Backless
sleeve: 3/4
color: Navy; Fabric: Jersey;
Pattern: print; Category: Shirt;
Fit: large; Gender: Female
Neckline: Square
Sleeve: Short
Caption: "Shirt is Navy color and Jersey fabric and Large fit and Backless neckline and Print pattern and 3/4 sleeve"
Query text: Shirt, replace Backless neckline with Square neckline, and replace 3/4 sleeve with Short sleeve"
Caption: "Shirt is Navy color and Jersey fabric and Large fit and Square neckline and Print pattern and Short sleeve"

Figure 3.5: Example of image pair and generated text query from Shopping100k dataset. Gray words indicate shared attributes.

ative margin of 9.4%. We also observe that token based methods, namely MAAF and AACL, perform better than residual based methods. This indicates that the rich information contained in tokens is beneficial for feature composition. Figure 3.4 shows our qualitative results on Fashion200k. Our model is able to retrieve new images that resemble the reference image, while changing certain attributes conditioned on text feedback—e.g., fit, color and length. We also observe that all retrieved images share the same semantics and are visually similar to the target image, indicating the quantitative performance is potentially underestimated.

3.5.4 Shopping100k

Shopping100k [2] is a large-scale fashion dataset of individual clothing items extracted from different e-commence providers. It contains 101,021 images of 12 fashion attributes, covering the following categories: “collar”, “color”, “fabric”, “fastening”, “fit”, “gender”, “length”, “neckline”, “pattern”, “pocket”, “sleeve length”, and “sport”. A total of 151 different labels are generated by combinations of different attributes and the corresponding attributes values. Compared to FashionIQ and Fashion200k, the Shopping100k dataset is more diverse and only contains garments in isolation. In addition, FashionIQ and Fashion200k only contain 3 and 5 apparel categories, respectively.

Each image in Shopping100k is tagged with the attributes and attribute values, such

Model	Dress	Jacket	Jean	Jumper	Shirt	Shorts	Skirt	Swimming	T-shirt	Bottoms	Trouser	Average
Recall@1												
TIRG*	6.81±0.58	10.46±0.97	4.83±1.43	11.87±1.26	13.15±1.25	12.38±1.16	10.92±1.22	13.51±1.49	11.87±0.80	8.32±0.60	13.03±1.77	10.65±0.37
MAAF*	7.05±0.86	12.43±0.76	5.79±1.34	13.19±0.88	**14.44±1.28**	13.21±1.68	12.11±0.77	12.41±0.71	12.89±1.16	**10.28±1.35**	12.89±0.87	11.52±0.39
RTIC*	6.80±0.09	11.70±0.90	5.27±0.90	12.08±1.39	13.93±1.33	11.83±0.97	10.96±1.44	13.18±0.99	12.60±0.99	8.49±0.65	11.70±1.70	10.78±0.44
AACL	**7.70±0.67**	**12.63±0.93**	**7.27±0.96**	**13.30±0.31**	14.21±0.52	**14.38±1.14**	**14.55±1.22**	**16.22±1.02**	**13.66±0.28**	10.00±0.53	**14.14±0.63**	**12.55±0.32**
Recall@10												
TIRG*	34.22±0.53	49.86±0.47	29.23±0.48	51.08±0.89	50.22±0.72	50.43±0.52	55.85±0.58	51.86±1.49	47.19±1.04	41.69±0.59	51.06±1.28	46.61±0.35
MAAF*	35.01±1.85	51.48±1.67	**31.78±1.12**	51.70±2.45	52.15±1.96	50.64±1.30	54.70±3.36	54.74±2.46	**49.31±1.79**	44.00±2.87	52.08±0.63	47.96±0.65
RTIC*	33.17±1.92	50.51±2.11	29.21±4.36	48.92±3.39	50.90±2.89	50.29±0.74	51.96±2.09	51.62±2.02	46.71±2.41	42.24±1.31	51.46±1.25	46.09±1.03
AACL	**35.16±0.54**	**51.63±1.33**	30.80±1.79	**52.31±0.89**	**52.52±1.32**	**54.63±1.66**	**57.54±0.95**	**56.13±2.13**	49.18±1.40	**46.69±1.06**	**54.63±1.72**	**49.20±0.46**
Recall@50												
TIRG*	66.15± 0.80	81.50±0.38	62.47±0.19	80.74±2.40	82.43±0.28	81.36±0.95	85.57±1.66	83.91±1.20	79.32±1.81	77.94±1.18	85.02±1.35	78.76± 0.69
MAAF*	68.42±1.42	82.73±2.29	63.24±2.94	82.28±1.36	84.41±1.90	82.06±1.66	88.19±0.78	**85.32±2.27**	**81.07±1.34**	81.17±0.67	86.75±0.82	80.51±0.56
RTIC*	67.30±2.12	81.92± 2.42	**64.30±5.31**	80.27±2.37	83.45±1.58	82.22±1.88	84.71±1.57	84.15±2.46	78.87±1.95	79.47±0.88	85.37±1.92	79.27±1.12
AACL	**69.21±0.37**	**83.30±1.77**	63.92±3.59	**82.30±0.36**	**84.75±1.21**	**85.50±1.30**	**88.94±0.78**	85.31±1.52	80.54±1.18	**82.83±0.88**	**87.61±0.76**	**81.29±1.11**

Table 3.4: Comparison of image search with text feedback on our modified Shopping100k dataset. Averages are computed over all categories. * denotes our implementation results obtained with the same image encoder and text encoder as AACL.

as *"Neckline: Backless, Sleeve: 3/4, Color: Navy, Fabric: Jersey, Pattern: Print, Category: Shirt, Fit: Large, Gender: Female"*. There are 15 high-level apparel categories. To generate the dataset for image retrieval with text feedback, we remove categories that contain fewer than 2,000 images, namely "coat", "suit", "jumpsuit", "pyjamas", and "tracksuit". The final set of 11 categories is listed in Table 3.3 along with the number of images in each category. A training split with 76,867 images and a validation split with 19,210 images is randomly sampled from these remaining categories.

To generate the training image pairs and modification text, we first derive a descriptive caption for each image using its tagged attribute values by concatenating the category with "is", followed by attributes joined by "and"—e.g., *"Shirt is Navy color and Jersey fabric and Large fit and Backless neckline and Print pattern and 3/4 sleeve"*. Queries are created by selecting image pairs that differ in two attributes in the description. Note that we constrain the image pairs to be from the same apparel category and gender. The modification text is created with the apparel category plus the attribute modifications following the pattern "replace xx with xx"—i.e. *"Shirt, replace Backless neckline with Square neckline, and replace 3/4 sleeve with Short sleeve."* (Figure 3.5). During training, the query and target image pairs are selected on-the-fly based on the number of attributes we specify. For our experiments, 16,237 fixed test query pairs are generated from the validation set for performance evaluation.

Table 3.4 compares our approach to other methods on Shopping100k. Our model

Stage(s)	Recall@10	Recall@50
Stage 2 + 3 + 4	48.78	80.74
Stage 3 + 4	**49.20**	**81.29**
Stage 4	48.56	81.25

Table 3.5: Ablation of using tokens from different Swin Transformer stages on our modified Shopping100k dataset.

is shown to clearly outperform the SOTA baselines. Figure 3.6 presents some qualitative examples. These examples yield three observations. First, our model is capable of understanding rich image-text representations, including global attributes such as color, pattern, and fit, as well as local attributes such as collar, neckline, and sleeves. Second, our model is capable of using the text information to selectively modify the query images. As an example, for the first query the retrieved images preserve the striped pattern even though it is not requested in the text feedback. Five of the top-5 retrieved candidates fulfill the "long sleeves" requirement and four candidates have "low-v-neck". Third, the model is capable of capturing minor modifications such as "kent collar" *vs.* "mandarin collar", suggesting it can be successfully utilized in fine-grained search.

3.5.5 Ablation Study

Image representation: Table 3.5 compares the performance of AACL when using different image representations from the Swin Transformer on our modified Shopping100k dataset. The experiments reveal that using image tokens from Stages 3 and 4 is most effective for this task. The concatenation of two stages from the encoder considers richer forms of image representation. Somewhat surprisingly, concatenating representations from Stage 2 does not seem to benefit the task. This may suggest that at some point, the lower level information may distract the model from capturing meaningful global contextual information.

Number of attributes: In Table 3.6, we see the effect of the number of attributes that differ on the Shopping100k dataset. We constrain the modification text to have varying numbers of differing attributes: 2 attributes, 1 or 2 attributes, or 1 attribute. Having 2 differing attributes is seen to be the most difficult case and so we choose it to compare with the other methods in Table 3.4.

Dataset	Dress	Jacket	Jean	Jumper	Shirt	Shorts	Skirt	Swimming	T-shirt	Bottoms	Trouser	Average
Recall@10												
1 attribute	30.11	52.48	47.71	53.34	51.36	55.39	52.17	54.29	48.90	47.69	50.24	49.42
1 or 2 attributes	33.64	54.25	46.28	54.38	51.11	54.36	55.25	58.98	52.09	46.44	49.43	50.56
2 attributes	35.16	51.63	30.80	52.31	52.52	54.63	57.54	56.13	49.18	46.69	54.63	49.20
Recall@50												
1 attribute	61.52	84.78	80.09	84.02	84.10	86.83	89.02	84.70	81.04	80.03	84.66	81.89
1 or 2 attributes	66.78	85.16	81.48	83.38	82.72	86.25	89.62	87.42	81.67	83.07	86.81	83.12
2 attributes	69.21	83.30	63.92	82.30	84.75	85.50	88.94	85.31	80.54	82.83	87.61	81.29

Table 3.6: Comparison of AACL when constructing Shopping100k dataset with different number of attributes. Lowest value is underlined.

Method	Recall@10	Recall@50
Additive→Dot-Product	48.37	80.14
Product→Addition	48.56	80.45
AACL	**49.20**	**81.29**

Table 3.7: Ablation of self-attention layer on our modified Shopping100k dataset. We separately examine substituting additive self-attention with standard dot-product and changing the Hadamard product to addition.

Additive attention:

To assess the importance of additive attention, we perform a comparison by substituting with dot-product attention. Table 3.7 “Additive→Dot-Product” shows the comparison on our modified Shopping100k dataset. From these results, we that AACL does benefit consistently from the additive attention. In addition, dot product attention is more computationally expensive than additive attention ($O(n^2)$ *vs.* $O(n)$) and as such the benefits of additive attention extend beyond evaluation performance gains.

Interaction function:

We study the effect of using different functions, namely addition and Hadamard product, to model the interactions between the context vector and the individual tokens. We compare the standard AACL and this variant on Shopping100k. The results are shown in Table 3.7 “Product→Addition”. The Hadamard product performs consistently better than addition, indicating this form of non-linear modeling is beneficial.

3.5.6 Additive Attention Visualization

To interpret the attention learned by AACL, we count the number of instances of words with high normalized attention scores from the FashionIQ validation set. The word attention scores are normalized as follows: We first multiply the α_i in Equation 3.2 across all blocks to get the total attention flow for each token. Subsequently, the minimum word token flow score is mapped to zero and the maximum to one. We apply a threshold of 0.8 for high scores.

To further interpret the attention learned by AACL, we visualize the attended regions in Figure 3.7. We apply a mask based on the attention flow (as calculated above) to the input query image. Note that, since we are using the Swin Transformer as the image encoder, the encoded feature maps are 7×7 and the resulting visualization resolution appears lower than with other models. Nevertheless, we do observe that the spatially attended regions vary with the query text. This indicates that the additive self-attention selects different visual content to transform conditioned on the text query.

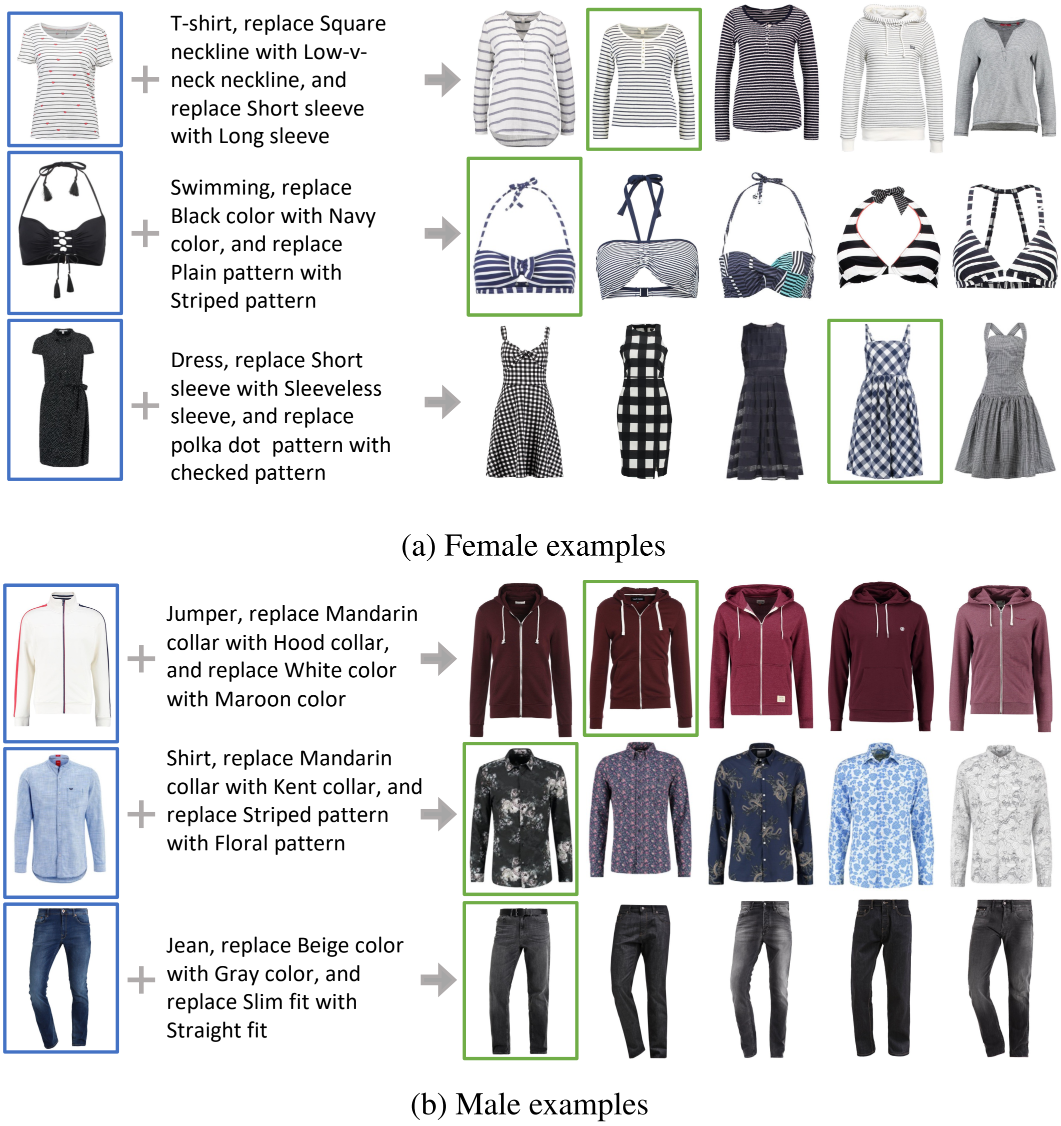
3.5.7 Additional Qualitative Results

We present additional qualitative results on the FashionIQ dataset to provide further insight into using natural language as text feedback. Note that the query text of FashionIQ most closely resembles natural language as the queries are provided by annotators from English-speaking countries.

Figure 3.8 qualitatively compares our AACL model with TIRG, RTIC and MAAF on the FashionIQ dataset Dress category. Figures 3.9 and 3.10 further illustrate the retrieval results on the Toptee and Shirt categories, respectively. We present the query image and query text in the first row, followed by the top-5 retrieved images from the various models in subsequent rows. Even though for each query image a single target image is defined, there can be multiple “perceptually acceptable” images. This is because there may exist multiple items in the database that are similar to the target image and satisfy the modifying text component of the query. In Figure 3.9b, for example, there is more than one toptee that is short sleeved with gray and white stripes among the retrieved items, but only the target image is considered a correct match. Compared to the other models considered, our AACL model tends to find the best matching images that satisfy all conditions in the queries.

3.6 Conclusion

We present AACL, a novel and general-purpose solution to the challenging task of image search with text feedback. This framework features an additive self-attention layer that selectively preserves and transforms multi-level visual features conditioned on text semantics to derive an expressive composite representation. We validate the efficacy of AACL on three datasets, and demonstrate its consistent superiority in handling various text feedback for natural language expression. Overall, our work provides a novel approach along with a comprehensive evaluation, which collectively advance the research in interactive visual search using text feedback.


T-shirt, replace Square neckline with Low-v-neck neckline, and replace Short sleeve with Long sleeve
Swimming, replace Black color with Navy color, and replace Plain pattern with Striped pattern
Dress, replace Short sleeve with Sleeveless sleeve, and replace polka dot pattern with checked pattern
Jumper, replace Mandarin collar with Hood collar, and replace White color with Maroon color
Shirt, replace Mandarin collar with Kent collar, and replace Striped pattern with Floral pattern
Jean, replace Beige color with Gray color, and replace Slim fit with Straight fit

(a) Female examples

(b) Male examples

Figure 3.6: Qualitative results of AACL on (a) Female and (b) Male set of our modified Shopping100k dataset. Blue and green box indicate query and target images, respectively.

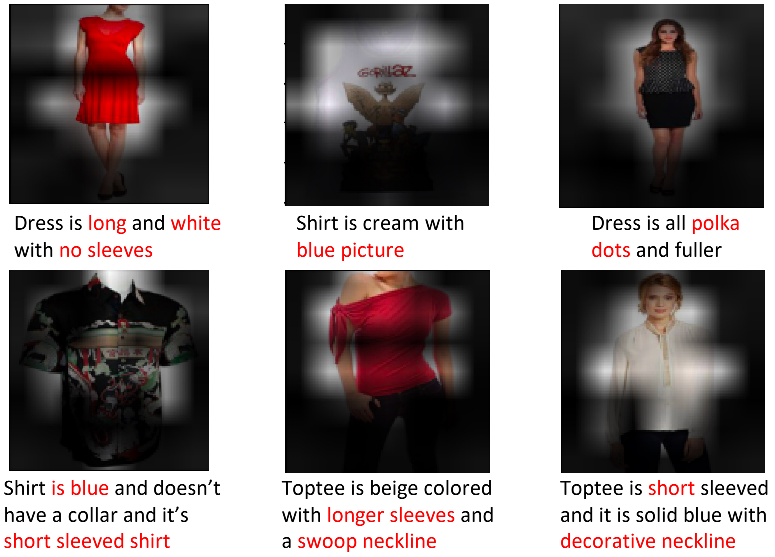
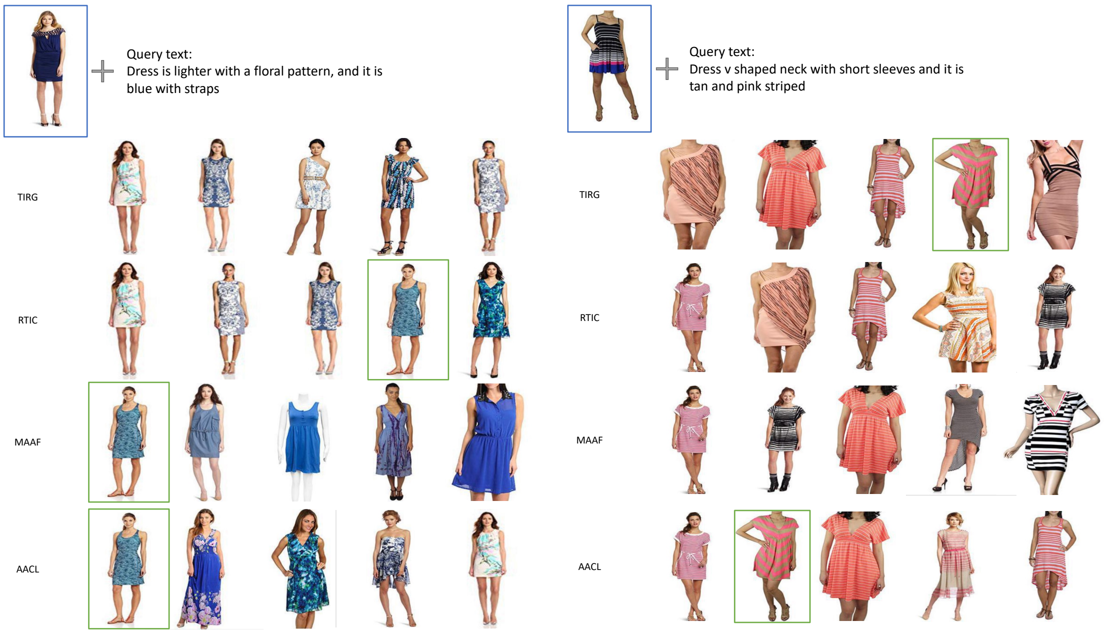
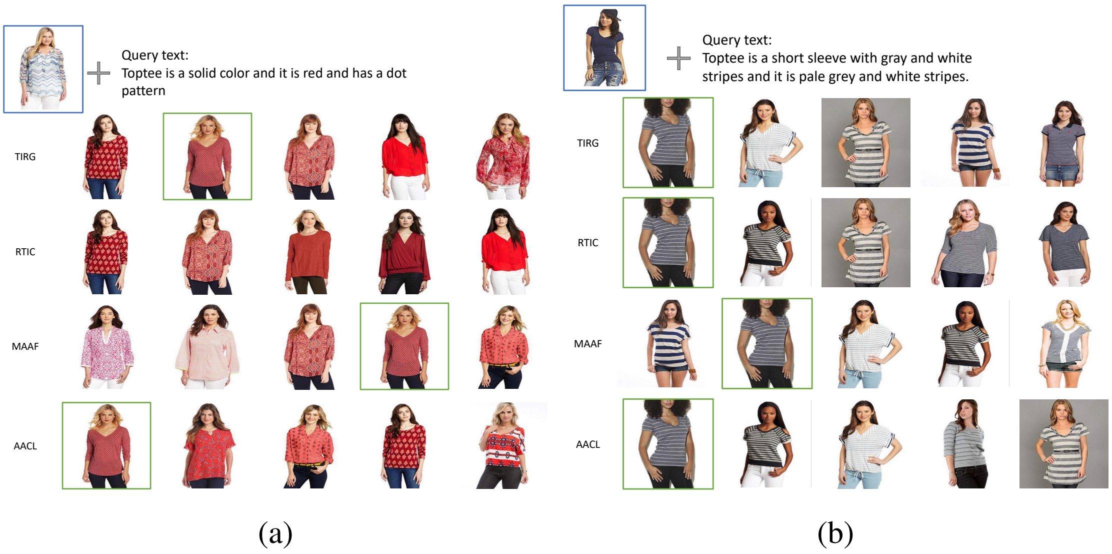

Dress is long and white with no sleeves
Shirt is cream with blue picture
Dress is all polka dots and fuller
Shirt is blue and doesn't have a collar and it's short sleeved shirt
Toptee is beige colored with longer sleeves and a swoop neckline
Toptee is short sleeved and it is solid blue with decorative neckline

Figure 3.7: Attention visualization of AACL model on FashionIQ dataset. Words with high attention value are in red.


Query text:
Dress is lighter with a floral pattern, and it is blue with straps
TIRG
RTIC
MAAF
AACL
Query text:
Dress v shaped neck with short sleeves and it is tan and pink striped
TIRG
RTIC
MAAF
AACL

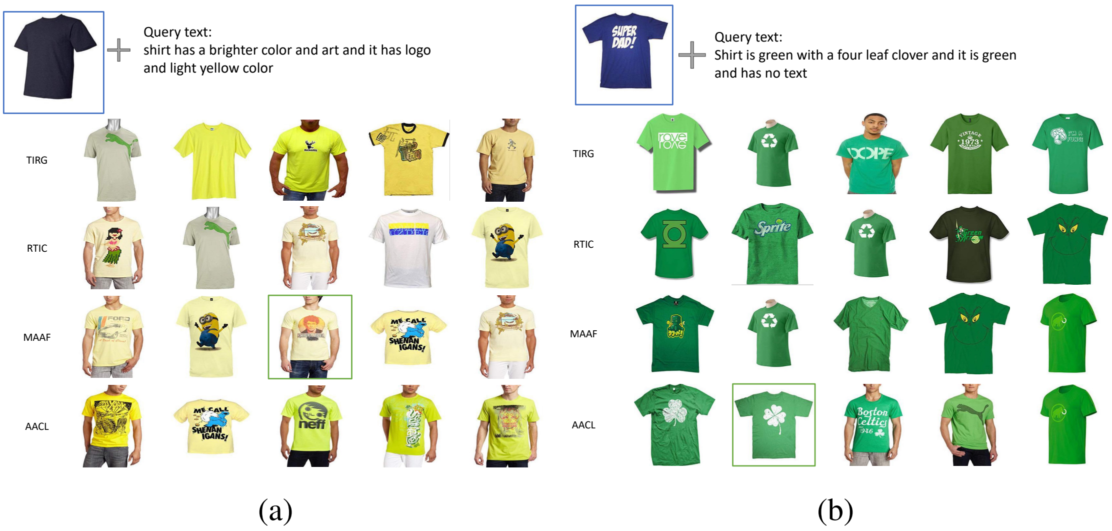
(a) (b)

Figure 3.8: Qualitative results on FashionIQ dataset Dress category. Blue and green box indicate query and target images, respectively.


Query text:
Toptee is a solid color and it is red and has a dot pattern
TIRG
RTIC
MAAF
AACL
Query text:
Toptee is a short sleeve with gray and white stripes and it is pale grey and white stripes.
TIRG
RTIC
MAAF
AACL

(a) (b)

Figure 3.9: Qualitative results on FashionIQ dataset Toptee category. Blue and green box indicate query and target images, respectively.


Query text:
shirt has a brighter color and art and it has logo and light yellow color
TIRG
RTIC
MAAF
AACL
Query text:
Shirt is green with a four leaf clover and it is green and has no text
TIRG
RTIC
MAAF
AACL

(a) (b)

Figure 3.10: Qualitative results on FashionIQ dataset Shirt category. Blue and green box indicate query and target images, respectively.

Chapter 4

Que2Engage: Embedding-based Retrieval for Relevant and Engaging Products at Facebook Marketplace

4.1 Overview

Embedding-Based Retrieval (EBR) in e-commerce search is a powerful search retrieval technique to address semantic matches between search queries and products. However, commercial search engines like Facebook Marketplace Search are complex multi-stage systems optimized for multiple business objectives. At Facebook Marketplace, search retrieval focuses on matching search queries with relevant products, while search ranking puts more emphasis on contextual signals to up-rank the more engaging products. As a result, the end-to-end searcher experience is a combination of both relevance and engagement, and the interaction between different stages of the system. This presents challenges to EBR systems in order to optimize for better searcher experiences. In this chapter we presents Que2Engage, a search EBR system built towards bridging the gap between retrieval and ranking for end-to-end optimizations. Que2Engage takes a multimodal and multitask approach to infuse contextual information into the retrieval stage and to balance different business objectives. We show the effectiveness of our approach via a multitask evaluation framework and thorough baseline comparisons and ablation studies. Que2Engage is deployed on Facebook Marketplace Search and shows significant improvements in searcher engagement in two weeks of A/B testing.

4.2 Introduction

Recent years have witnessed surprising advances in machine learning (ML), which in turn have led to the pervasive application of ML models across several domains [22, 21, 78, 81]. Embedding-based Retrieval (EBR) has become an important component of e-commerce search engines across Facebook Marketplace, Walmart, Instacart, and more [66, 73, 83, 128, 44]. In general, EBR models focus on learning embedding representations for search queries and documents, so that documents semantically close to a search query can be retrieved via ANN search [66, 73, 83, 34]. However, search engines are usually complex multi-stage systems optimized for multiple business objectives, so simply optimizing for semantic relevance may not always lead to the best outcome. For example, [73] points out that integrating EBR systems can lead to Normalized Discounted Cumulative Gain (NDCG) regressions because downstream re-ranking systems may not always able to rank results retrieved via EBR properly.

In e-commerce platforms like Facebook Marketplace [1], contextual information such as product price, production condition, seller rating, *etc.* are also important signals to consider to ensure products retrieved are engaging to the searchers. However, we argue that the leverage of contextual information in a search EBR setting towards better searcher engagement is not a trivial problem because (1) traditional EBR modeling techniques based on contrastive learning overly emphasize on semantic relevance, so naively applying contextual information in a contrastive learning setting may not work well (2) a product being semantically relevant to a query does not imply that it is engaging to the searcher, and thus simultaneously preserving relevance and engagement is challenging.

In this chapter we present Que2Engage, an extension of Que2Search [73] to addresses the aforementioned challenges. It takes a multimodal approach to incorporate contextual signals as a unique modality in its transformer fusion backbone. The model is trained with multitask learning that joins contrastive learning with ranker-style training to not only retrieve semantically relevant products, but also up-ranks the more engaging products like a re-ranking model. Similar to [137], we propose a multitask evaluation for EBR models to understand its performances in different domains. We share detailed baseline comparisons and ablation studies using the multitask evaluation framework to illustrate our argument of multi-stage consistency and the effectiveness of our approach in leveraging contextual information.

Que2Engage is integrated in Facebook Marketplace Search and powering millions of search queries per day. It has demonstrated significant improvements in searcher engagement via two weeks of online A/B testing.

4.3 Related Work

4.3.1 Embedding-based Retrieval

The largest E-commerce sites offer over millions of products for sale. Choosing among so many options is challenging for both individual and group-oriented cus-

[1] www.facebook.com/marketplace

tomers [139]. To alleviate the information overload caused by the tremendous amount of data that existing online services expose to end-users, recommendation systems have emerged to help customers choose the best “content” [53]. A recommendation system for an e-commerce site receives information from a consumer about the search intent and recommends products that are likely to meet needs. Today, recommendation systems are deployed on hundreds of different sites, serving millions of consumers. In recent years, Embedding-Based Retrieval (EBR) has been adopted in e-commerce search to retrieve semantically relevant products as a complement of lexical retrieval [73, 83, 66, 137]. Siamese neural networks [15, 105] trained with contrastive learning loss [125, 106, 107] are among the popular modeling choices for EBR in both search and recommendation systems [112, 129, 77].

4.3.2 Contrastive Learning

Training recommendation systems on large item databases often involves treating the process as an extreme multi-class classification task, where negative sampling is crucial. The common method for two-tower models involves using in-batch negatives—positive items of other users within the same mini-batch are considered as negative items. This method has become a standard practice for reducing computational load and increasing training efficiency [112]. However, naive contrastive learning using in-batch negatives can suffer from missing interesting negative samples [137, 129] and being memory-hungry [112]. Variants of contrastive learnings are proposed to address them by incorporating smarter negative sampling [129, 137, 73] and optimizing memory usage [112, 137]. Sometimes, teacher-student learning is also used as an auxiliary task to improve relevance [34, 137].

4.3.3 Contextual Information

In search retrieval, Pre-trained Language Models (PLM) are widely adopted because the main focus of retrieval is often textual relevance [83, 77]. The exploration of personalization models, data knowledge, and the generation of appropriate product recommendations to enhance user engagement has recently gained interest from the research

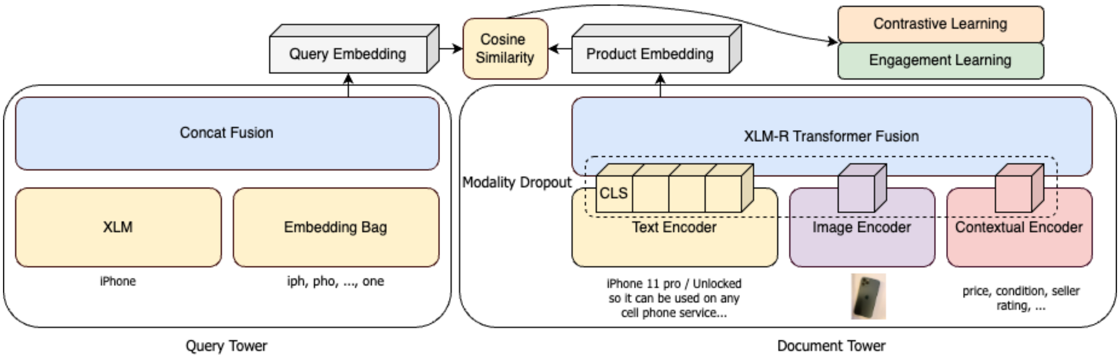


Figure 4.1: Que2Engage architecture overview

community [59, 82]. Research [66, 73, 137] points out that contextual information and consistency with re-ranking stage are also important factors to EBR systems' end-to-end performance that are beyond textual relevance. Recently, [137] developed a unified training scheme to balance multiple optimization objectives, yet the role of contextual information in the multi-objective setting is rarely discussed.

4.4 Modeling

Figure 4.1 presents the overall architecture of our Que2Engage framework, which is a two-tower neural network consisting of a query and a document tower for learning embedding representations of search queries and e-commerce products, respectively. Multimodal and contextual information of products are fed into the document tower using a transformer-fusion approach, and trained with multi-task learning. In the following, we detail our choices in model architecture and our novel multimodal multitask method.

4.4.1 Model Architecture

Query tower

Similar to [73], we adopt a multi-granular representation of search queries which consists of both raw query text and character trigrams of the query. Raw query text is encoded using a 2-layer XLM [24] encoder, and character trigrams are encoded using an

EmbeddingBag [86] encoder. Different from [73], we combine the two representations using concatenation instead of attention fusion before sending to the final MLP layer as we find the former yields slightly better performance.

Contextual information as a modality

For candidate product listings, we use an MLP-based encoder from the document tower to encode the contextual information such as price, category and creation time. Numerical features are represented as single neurons, and categorical features are represented using one-hot encoding. All of the contextual features are concatenated and then fed into a BatchNorm layer followed by a final MLP to ensure a fixed numerical scale and a fixed output length. We call this encoder output a ”context token” because it is treated similarly to text and image tokens during the multimodal fusion step covered in section 4.4.1. Essentially, contextual information is treated as a unique modality in our multimodal framework.

Multimodal fusion

Besides encoding the contextual information, we use text encoder to convert the textual fields, *i.e.*, product title and description, into a sequence of word tokens and feed them into the transformer to get the textual embedding. A special [CLS] token is used to encode the whole sentence representation. For the variable number of images attached to the document, we take the pre-trained image representations [12] for each of the attached images, apply a shared MLP layer and deep sets [134] fusion to get the image dense representation as an image modality token. We borrow the transformer-fusion architecture used in [132], where we feed the concatenation of the text tokens, image token as well as context token to the multimodal fusion encoder. Our text encoder and multimodal fusion encoder are initialized from 6-layer XLM-R [23], an multilingual language model. As in [132], the text encoder inherits its first K layers and and the multimodal fusion model inherits its remaining M layers. We extract the hidden output of the [CLS] token at the last layer of multimodal fusion encoder and project it to the desired dimension as the final document embedding.

Modality dropout

To ensure that our model does not overly rely on one modality and is robust against missing information during inference time, we introduce a modality dropout mechanism. Specifically, we randomly mask out the output of contextual encoder, image encoder and text encoder with probability of δ_c, δ_i and δ_t respectively, and replace the masked ones with tensor of zero.

Learning image representation

We explore two variants of image encoders to capture visual information from product images. In method one, we directly apply pre-computed image embedding from the GrokNet model [12], with an MLP layer on top to ensure the image embedding size is consistent with all the other tokens in the transformer fusion. In method two, we include an off-the-shelf RestNet50 encoder from the CommerceMM model [132] into the document tower, and train the entire document tower as a continued CommerceMM fine-tuning process. While the latter is obviously more powerful because the original image encoder is retained and fine-tuned, the former is much simpler to train and requires less memory in both training and serving. We will compare them in more detail in section 4.5.

4.4.2 Multitask Training

Contrastive learning

We adopt contrastive learning based on batch negative sampling as part of our training objectives, where positive samples are user engaged $\langle$ query, product $\rangle$ pairs sampled from anonymized search logs, and negative samples are generated by randomly combining queries and products within a mini-batch of positive samples. Formally, we introduce the relevance loss $L_{relevance}$ as follows

$$L_{relevance} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{s \cdot \kappa\left(q_i, d_i\right)\right\}}{\sum_{j=1}^{B}\exp\left\{s \cdot \kappa\left(q_i, d_j\right)\right\}}\right\} \tag{4.1}$$

where B is the batch size, κ is a similarity kernel that is implemented as the cosine similarity, and s denotes a scaling factor which is simply a fixed value $s = 20$ throughout all experiments.

Learning contextual information

Although batch negative sampling has proven useful in learning semantic relevance in the search EBR problem [83, 73], we notice that it is not sufficient in learning contextual information towards user engagement. For example, contextual information like product price is a distinguishing factor to identify engaging products among relevant products (*i.e.* a product can be relevant but receives no engagement because the listed price is not reasonable). However, during batch negative training, the model receives little negative supervision from products with very unreasonable prices, since all negative samples are generated from engaged $\langle$ query, product $\rangle$ pairs. Fundamentally, as [125] points out, this is because batch negative methods implicitly sample from the distribution of engaged products, which may not be the true distribution of the inventory. Methods like mixing random negatives [129] and in-batch hard negative mining [73] are proposed to mitigate the problem. However, we observe that negative samples generated by those approaches are still too easy for the model to pick up the nuances in contextual information. Therefore, we propose an auxiliary training task that optimizes the model directly towards finding engaging products among relevant products. Specifically, we augment the training set in section 4.4.2 by including $\langle$ query, product $\rangle$ pairs displayed to the searchers but which receive no searcher engagements as hard negatives, and compute a BCE loss on those samples. Formally, we define the loss $L_{engagement}$ as follows

$$L_{engagement} = -(y_i \log(c_i) + (1 - y_i) \log(1 - c_i)) \tag{4.2}$$

where $c_i = s \cdot \kappa\,(q_i, d_i)$.

To combine $L_{relevance}$ and $L_{engagement}$, we define the final multitask loss as

$$L(\theta) = \lambda_1 \cdot L_{relevance} + \lambda_2 \cdot L_{engagement} \tag{4.3}$$

where θ is the model parameters, λ_1 and λ_2 are the weighting parameters chosen empirically.

4.5 Offline Experiments

4.5.1 Data Collections

We collect 150 million ⟨ query, product ⟩ pairs displayed to searchers from Facebook Marketplace's search log. To avoid the potential imbalance issue [61], we sample 75 million pairs receiving downstream engagements as positive samples, and 75 million pairs as negative samples. The data is de-identified and aggregated before evaluation proceeds. For offline evaluation, we collect 26k human-rated data as the relevance evaluation set. The human-rated dataset is generated by letting raters to decide whether a result is relevant to a query or not. The candidates to be rated are generated by a stratified sampling of the search queries and products, which includes both easy and hard samples. To evaluate user engagement, we reserve one future date among the 150 million de-identified and aggregated search log data.

4.5.2 Baselines and Ablation Studies

We choose Que2Search [73] as our baseline model, which is a two-tower model based on attention fusion of pre-trained XLM encoders [24] and image representations. We further augment the model with encoders based on contextual information, as well as with mixed batch method [129] to incorporate hard negatives from products displayed to searchers. For the treatment group, we use the Que2Engage with pre-computed image embeddings, because in practice its simplicity is preferred during the actual model productionalization, and we share the comparison against an alternative image encoder based on fine-tuning of the CommerceMM encoder separately.

4.5.3 Experimental Setup

Evaluation metrics

Similar to [137], we adopt a multitask evaluation framework to measure semantic relevance and searcher engagement separately. Semantic relevance is measured using the 26k human-rated dataset, and searcher engagement is measured using the 220k future engagement dataset. For both datasets, we rank them using the cosine similarity from our model and report ROC_AUC as the evaluation metric. Note that one significant difference between the two datasets lies in personalization. For the human-rated dataset, raters are asked to make judgement based purely on an objective guideline around textual and visual relevance (*e.g.* whether there is a catalog or brand mismatch between a query and a product), while more subjective factors outside of the guideline (*e.g.* whether the listed price or product condition is appealing) play important roles in the engagement dataset.

Therefore, ROC_AUC on the relevance dataset measures how well the model predicts search relevance (similar to the in relevance degree in [137]), which is the main evaluation metric used in [73]. We have also found it correlate well with search retrieval performance. ROC_AUC of the engagement dataset essentially measures its performance on the search ranking task, because the candidates are all products with user impressions. In fact, it is also a good indicator of the consistency between search retrieval and search ranking.

Experimental parameters

We develop all of the models on Nvidia A100 GPUs using the PyTorch Multimodal framework [102]. Models are trained using batch size of 512, and optimized using Adam optimizer with a learning rate of $4e-4$. We set weighting parameters λ_1 and λ_2 as 0.8 and 0.2 respectively. For the modality dropout, δ_c, δ_i and δ_t are 0.5, 0 and 0.5. We directly feed the text tokens and the tokens from other modalities into the multimodal transformer as [132], *i.e.*, our multimodal transformer is an early-fusion model with 0-layer text encoder and 6-layer multimodal fusion encoder. One exception to the aforementioned settings is that when comparing pre-computed image embeddings with fine-tuning the CommerceMM encoder, due to the increased GPU memory consumption of the fine-tuning approach, we adjust the batch size to 64 for that particular experiment. The learning rate was also adjusted to $5e-5$ for the CommerceMM training to avoid NaN in the loss function computation.

4.5.4 Results

Analysis of baseline results

The top part of table 4.1 outlines the performance of baseline models. We can see that Que2Search does well on the relevance evaluation but performs poorly on the engagement dataset, which is expected because it is optimized for semantic relevance. While adding contextual information to Que2Search itself does not improve the performance on engagement evaluation, changing the training objective to include mixed negatives significantly improves the prediction. Mixed negatives also helps Que2Search to better leverage contextual features towards engagement prediction with row 4 in table 4.1 achieving the highest ROC_AUC on engagement evaluation. This aligns with our hypothesis in section 4.4.2 - vanilla batch negative is insufficient to learning the nuances in engagement prediction from contextual information due to the sampling bias introduced by generating negatives from the positives, and this can be mitigated by introducing mixed negatives. Note that rows 4 and 5 in table 4.1 suggest a regression in relevance evaluation along with the improvement of engagement evaluation. This is expected because contextual information like product price and condition are irrelevant to the query-product relevance guideline provided to our raters, and thus we do not expect the leverage of contextual signals to improve the relevance evaluation.

Que2Engage and ablation studies

The second part of table 4.1 shows that engagement evaluation can be significantly improved with the Que2Engage approach, with the full Que2Engage using multitask learning and modality dropout achieving the best results across the two evaluation methods. An ablation study on the loss function suggests that multitask training leads to the biggest improvement in engagement evaluation, suggesting that a more focused loss function on hard negatives works better than simply mixing the negatives. Finally, modality dropout further improves both metrics and specially the relevance evaluation, suggesting that forcing missing modalities may prevent the model over-fitting on one task and thus regressing the other.

Model	Engagement	Relevance
Que2Search[73]	55.88	67.14
Que2Search w/ contextual encoder	55.85	66.74
Que2Search w/ mixed batch loss	63.63	63.79
Que2Search w/ contextual encoder + mixed batch loss	64.45	61.17
Que2Engage w/ mixed batch loss	64.70	60.36
Que2Engage w/ multitask training	76.13	65.63
Que2Engage w/ multitask training + modality dropout	**76.90**	**67.21**

Table 4.1: Results for baseline comparison and ablation studies

Method	Engagement	Relevance
pre-computed image embedding	74.35	60.19
fine-tuned CommerceMM encoder	**74.67**	**60.55**

Table 4.2: Results for image encoder comparison

Image encoders

We also compare the two approaches to incorporate image signals into the multimodal fusion framework. We can see that although both methods are based on pretraining tasks, being able to fine-tune the image encoder indeed outperforms the original frozen GrokNet [12] approach used in Que2Search in both evaluations. However, given that fine-tuning the image encoder end-to-end requires significantly more GPU memory and training time, we do not include this technique in the production model for simplicity. And the ablation study is done with a reduced batch size of 64 which significantly regresses the absolute relevance evaluation because larger batch sizes have proven helpful in contrastive learning [112]. However, we hope to productionalize this technique one day with smarter memory efficiency optimizations such in [112, 137] and with more powerful hardware.

4.6 Online Experiments

We deploy Que2Engage on Facebook Marketplace Search as a parallel retrieval source to the traditional lexical-based search retrieval. For online A/B testing, we compare Que2Engage with Que2Search [73], which is our previous production model solely optimized for semantic relevance. We measure both NDCG and searcher engagement for this A/B testing. Note that NDCG is calculated from human-rated labels using simulated search results similar to how the human-rated evaluation set is generated. Two weeks of online A/B testing shows that Que2Engage improves online searcher engagement by 4.5% while keeping NDCG neutral, which aligns with our offline multitask evaluation results.

4.7 Conclusion

We present the need of EBR modeling to balance relevance and engagement in the real-world applications like Facebook Marketplace, and introduce Que2Engage, our latest search EBR system, to address the challenges. Through baseline comparisons and ablations studies, we show the effectiveness of our innovations in incorporating contextual signals, multimodal techniques, representation learning and multitask learning. We have deployed Que2Engage on Facebook Marketplace Search. Through two weeks of A/B testing, we show that it outperforms our existing state-of-the-art search EBR system [73] and significantly improved searcher engagement on product listed at Facebook Marketplace.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this dissertation, an innovative framework is first developed for learning invariant representations for overhead image geolocalization. Subsequently, I develop frameworks to learn multimodal representations not only for image search with text feedback but also for embedding-based retrieval in recommendation systems. More specifically:

- I introduced a novel method for learning invariant representations for cross-time and orientation-invariant overhead image geolocalization. This innovative approach allows for accurate and efficient large-scale geolocalization, remaining robust even when the query and reference images originate from different times or orientations.
- I then proposed a framework for learning multimodal representations using fashion imagery. The fusion of image and text data in the model enhances the search experience, offering state-of-the-art performance in image retrieval with text feedback.
- I also presented a process for optimizing multimodal representation during the retrieval phase within a complex recommendation system. The model effectively incorporates images, text, and contextual information, achieving a balance between relevance and user engagement.

The research conducted in this dissertation not only contributes to the current state of knowledge in the fields of computer vision and natural language processing, but also offers practical solutions for real-world applications in geolocalization, fashion retrieval, and recommendation systems.

5.2 Future Work

Looking ahead, there are several promising directions for future research. For instance, a potential direction could involve the integration of multi-modal data into geolocalization, as discussed in Chapter 2. The use of geographic location may enable the connection of various modalities and models, such as ground conditions and user reviews, fostering the development of more comprehensive and informative representations.

The availability of sufficient training data is paramount for deep learning. Unfortunately, such circumstances are rare in many real-world applications. For instance, while we can readily obtain images and associated descriptions from social media, acquiring relative captions given image pairs poses a significant challenge. Therefore, it becomes crucial to explore ways to maximize the utility of existing data [58, 33, 31, 32] and denoising the data [80]. In future work, I aim to develop a relative change captioning framework as an extension to the work presented in Chapter 3. This approach, rather than adhering to a specific pattern, seeks to describe image differences through natural language descriptions. Such training data could enhance the model's ability to comprehend human feedback and bolster the performance of image search with text feedback. Another promising follow-up work in Chapter 3 could be interleaved image and text generation. The generation model, diverging from solely focusing on the retrieval of the target image from our dataset, can yield more innovative results even when the desired outcome isn't present in the dataset. The generated text can stimulate additional interaction and gather more user feedback, while the generation model such as the stable diffusion model could be deployed to create images in response to this multi-round text feedback. The recent advancement in prompt tuning can be leveraged for satisfying performance on the multi-round generation [116, 115].

Various strategies have been suggested to circumvent the memory constraints during the deployment phase of deep learning models [119, 120, 118]. However, the necessity of reducing the model size from the design phase itself is crucial. Techniques such as pruning [9, 10], parallelization [11], and sparse training [8] with efficient strategies offer theoretical guarantees and the potential to significantly reduce computational, memory, and communication costs involved in training and inference of deep learning models. Our objective is to learn representations that strike a balance between resource efficiency and model effectiveness, potentially through grounding the pre-trained large language model representations to the visual domain to enable cross-modality interactions. This becomes particularly critical given the success of large language models, which unfortunately comes with a substantial computational cost.

In conclusion, deep representation learning serves as a fundamental component in both computer vision and natural language processing fields. The process of learning representations and establishing correspondences across different modalities holds significant potential.

Bibliography

[1] Kenan E. Ak, Ashraf A. Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Kenan E. Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A. Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[3] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[4] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[5] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition". In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *arXiv*, 2014.

[7] Mayank Bansal, Kostas Daniilidis, and Harpreet Sawhney. Ultra-wide Baseline Facade Matching for Geo-localization. In *European Conference on Computer Vision*, 2012.

[8] Runxue Bao, Bin Gu, and Heng Huang. Efficient approximate solution path algorithm for order weight l_1-norm with accuracy guarantee. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 958–963. IEEE, 2019.

[9] Runxue Bao, Bin Gu, and Heng Huang. Fast oscar and owl regression via safe screening rules. In *International Conference on Machine Learning*, pages 653–663. PMLR, 2020.

[10] Runxue Bao, Bin Gu, and Heng Huang. An accelerated doubly stochastic gradient method with faster explicit model identification. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 57–66, 2022.

[11] Runxue Bao, Xidong Wu, Wenhan Xian, and Heng Huang. Doubly sparse asynchronous learning for stochastic composite optimization. In *IJCAI*, 2022.

[12] Sean Bell, Yiqun Liu, Sami Alsheikh, Yina Tang, Edward Pizzi, M. Henning, Karun Singh, Omkar Parkhi, and Fedor Borisyuk. Groknet: Unified computer vision model trunk and embeddings for commerce. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining*, KDD '20, page 2608–2616, New York, NY, USA, 2020. Association for Computing Machinery.

[13] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] Ayan Kumar Bhunia, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a ”siamese” time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[16] Remi Cadene, Hedi Ben-Younes, Nicolas Thome, and Matthieu Cord. Murel: Multimodal Relational Reasoning for Visual Question Answering. In *CVPR*, 2019.

[17] Francesco Castaldo, Amir Roshan Zamir, Roland Angst, Francesco A. N. Palmieri, and Silvio Savarese. Semantic Cross-View Matching. In *IEEE International Conference on Computer Vision Workshop*, 2015.

[18] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Proceedings of the european conference on computer vision (ECCV)*, 2020.

[20] 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)*, 2020.

[21] Ziheng Chen, Fabrizio Silvestri, Gabriele Tolomei, Jia Wang, He Zhu, and Hongshik Ahn. Explain the explainer: Interpreting model-agnostic counterfactual explanations of a deep reinforcement learning agent. *IEEE Transactions on Artificial Intelligence*, pages 1–15, 2022.

[22] Ziheng Chen, Fabrizio Silvestri, Jia Wang, He Zhu, Hongshik Ahn, and Gabriele Tolomei. Relax: Reinforcement learning agent explainer for arbitrary predictive models. In *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*, CIKM '22, page 252–261, New York, NY, USA, 2022. Association for Computing Machinery.

[23] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[24] Alexis Conneau and Guillaume Lample. *Cross-Lingual Language Model Pretraining*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[25] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[26] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable Convolutional Networks. In *IEEE International Conference on Computer Vision*, 2017.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[28] 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.

[29] Mehul Divecha and Shawn Newsam. Large-scale Geolocalization of Overhead Imagery. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016.

[30] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020.

[31] Bo Dong, Yang Gao, Swarup Chandra, and Latifur Khan. Multistream classification with relative density ratio estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3478–3485, 2019.

[32] Bo Dong, Jinghui Guo, Zhuoyi Wang, Rong Wu, Yang Gao, and Latifur Khan. Regression prediction for geolocation aware through relative density ratio estimation. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1644–1649, 2019.

[33] Bo Dong, Yiyi Wang, Hanbo Sun, Yunji Wang, Alireza Hashemi, and Zheng Du. CML: A contrastive meta learning method to estimate human label confidence scores and reduce data collection cost. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 35–43, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[34] Miao Fan, Jiacheng Guo, Shuai Zhu, Shuo Miao, Mingming Sun, and Ping Li. Mobius: Towards the next generation of query-ad matching in baidu's sponsored search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining*, KDD '19, page 2509–2517, New York, NY, USA, 2019. Association for Computing Machinery.

[35] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-Supervised Representation Learning by Rotation Feature Decoupling. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[36] Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[37] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*, 2018.

[38] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven J. Rennie, and Rogério Schmidt Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019.

[40] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[41] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[42] James Hays and Alexei A. Efros. IM2GPS: Estimating Geographic Information from a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] Yunzhong He, Cong Zhang, Ruoyan Kong, Chaitanya Kulkarni, Qing Liu, Ashish Gandhe, Amit Nithianandan, and Arul Prakash. Hiercat: Hierarchical query categorization from weakly supervised data at facebook marketplace. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 331–335, New York, NY, USA, 2023. Association for Computing Machinery.

[45] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[47] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. Embedding-based retrieval in facebook search. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining*, Aug 2020.

[48] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[49] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*, 2015.

[50] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: Multimodal, multitask retrieval across languages, 2021.

[51] Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy. Sac: Semantic attention composition for text-conditioned image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[52] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[53] Ruoming Jin, Dong Li, Jing Gao, Zhi Liu, Li Chen, and Yang Zhou. Towards a better understanding of linear models for recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining*, KDD '21, page 776–785, New York, NY, USA, 2021. Association for Computing Machinery.

[54] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, 2016.

[55] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual Compositional Learning in Interactive Image Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence)*, 2021.

[56] Jongseok Kim, Youngjae Yu, Seunghwan Lee, and GunheeKim. Cycled compositional learning between images and text. In *arXiv*, 2021.

[57] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[58] Ruoyan Kong, Zhanlong Qiu, Yang Liu, and Qi Zhao. Nimblelearn: A scalable and fast batch-mode active learning approach. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 350–359, 2021.

[59] Ruoyan Kong, Charles Chuankai Zhang, Ruixuan Sun, Vishnu Chhabra, Tanushsrisai Nadimpalli, and Joseph A. Konstan. Multi-objective personalization in multi-stakeholder organizational bulk e-mail. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27, Nov 2022.

[60] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 2015.

[61] Zhengfeng Lai, Chao Wang, Sen-ching Cheung, and Chen-Nee Chuah. Sar: Self-adaptive refinement on pseudo labels for multiclass-imbalanced semi-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[62] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. In *Nature*, 2015.

[63] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[64] Dong Li, Yelong Shen, Ruoming Jin, Yi Mao, Kuan Wang, and Weizhu Chen. Generation-augmented query expansion for code retrieval, 2022.

[65] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[66] Sen Li, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. Embedding-based product retrieval in taobao search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &amp; Data Mining*, KDD '21, page 3181–3189, New York, NY, USA, 2021. Association for Computing Machinery.

[67] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-View Image Geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[68] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[69] Tsung-Yi Lin, Yin Cui, Serge J. Belongie, and James Hays. Learning Deep Representations for Ground-to-aerial Geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[70] Yen-Liang Lin, Son Tran, and Larry S. Davis. Fashion outfit complementary item retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[71] L. Liu, H. Li, and Y. Dai. Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map. In *2017 IEEE International Conference on Computer Vision*, 2017.

[72] Liu Liu and Hongdong Li. Lending Orientation to Neural Networks for Cross-view Geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[73] Yiqun Liu, Kaushik Rangadurai, Yunzhong He, Siddarth Malreddy, Xunlong Gui, Xiaoyi Liu, and Fedor Borisyuk. Que2search: Fast and accurate query and document understanding for search at facebook. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &amp; Data Mining*, KDD '21, page 3376–3384, New York, NY, USA, 2021. Association for Computing Machinery.

[74] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[75] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. In *International Journal of Computer Vision*, 2004.

[76] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019.

[77] Wenhao Lu, Jian Jiao, and Ruofei Zhang. Twinbert: Distilling knowledge to twin-structured compressed bert models for large-scale retrieval. In *Proceedings of the 29th ACM International Conference on Information &amp; Knowledge Management*, CIKM '20, page 2645–2652, New York, NY, USA, 2020. Association for Computing Machinery.

[78] Xiaoling Luo, Xiaobo Ma, Matthew Munden, Yao-Jan Wu, and Yangsheng Jiang. A multisource data approach for estimating vehicle queue length at metered on-ramps. *Journal of Transportation Engineering, Part A: Systems*, page 04021117, 2022.

[79] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[80] He Lyu, Ningyu Sha, Shuyang Qin, Ming Yan, Yuying Xie, and Rongrong Wang. Manifold denoising by nonlinear robust principal component analysis. *Advances in neural information processing systems*, 32, 2019.

[81] Xiaobo Ma. Traffic performance evaluation using statistical and machine learning methods. *The University of Arizona*, 2022.

[82] Xiaobo Ma, Abolfazl Karimpour, and Yao-Jan Wu. Statistical evaluation of data requirement for ramp metering performance assessment. *Transportation Research Part A: Policy and Practice*, 2020.

[83] Alessandro Magnani, Feng Liu, Suthee Chaidaroon, Sachin Yadav, Praveen Reddy Suram, Ajit Puthenputhussery, Sijie Chen, Min Xie, Anirudh Kashi, Tony Lee, and Ciya Liao. Semantic retrieval at walmart. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3495–3503, New York, NY, USA, 2022. Association for Computing Machinery.

[84] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[85] Devi Parikh and Kristen Grauman. Relative attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2011.

[86] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[87] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence)*, 2018.

[88] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[89] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[90] F. Radenović, G. Tolias, and O. Chum. Fine-Tuning CNN Image Retrieval with No Human Annotation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[91] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[92] Abu Hasnat Mohammad Rubaiyat, Yongming Qin, and Homa Alemzadeh. Experimental resilience assessment of an open-source driving agent. *2018 IEEE 23rd Pacific Rim International Symposium on Dependable Computing (PRDC)*, pages 54–63, 2018.

[93] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[94] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[95] T. Sattler, B. Leibe, and L. Kobbelt. Fast Image-based Localization Using Direct 2D-to-3D Matching. In *International Conference on Computer Vision*, 2011.

[96] Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M. Seitz. Accurate Geo-Registration by Ground-to-Aerial Image Matching. In *International Conference on 3D Vision*, 2014.

[97] Yujiao Shi, Xin Yu Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal Feature Transport for Cross-View Image Geo-Localization. In *ArXiv*, 2019.

[98] Raymond Shiau, Hao-Yu Wu, Eric Kim, Yue Li Du, Anqi Guo, Zhiyuan Zhang, Eileen Li, Kunlong Gu, Charles Rosenberg, and Andrew Zhai. Shop the look. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[99] Minchul Shin, Yoonjae Cho, and Seongwuk Hong. Fashion-iq 2020 challenge 2nd place team's solution, 2020.

[100] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. Rtic: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*, 2021.

[101] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *IEEE International Conference on Learning Representations*, 2015.

[102] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research, 2020.

[103] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[104] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-View Image Matching for Geo-Localization in Urban Environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[105] Yuxin Tian, Xueqing Deng, Yi Zhu, and Shawn Newsam. Cross-time and orientation-invariant overhead image geolocalization using deep local features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[106] Yuxin Tian, Shawn Newsam, and Kofi Boakye. Image search with text feedback by additive attention compositional learning, 2022.

[107] Yuxin Tian, Shawn Newsam, and Kofi Boakye. Fashion image retrieval with text feedback by additive attention compositional learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1011–1021, January 2023.

[108] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[109] 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*, 2017.

[110] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[111] Nam N. Vo and James Hays. Localizing and Orienting Street Views Using Overhead Imagery. In *European Conference on Computer Vision*, 2016.

[112] Jinpeng Wang, Jieming Zhu, and Xiuqiang He. Cross-batch negative sampling for training two-tower recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1632–1636, New York, NY, USA, 2021. Association for Computing Machinery.

[113] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. Item silk road. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.

[114] Zhuoyi Wang, Yigong Wang, Bo Dong, Sahoo Pracheta, Kevin Hamlen, and Latifur Khan. Adaptive margin based deep adversarial metric learning. In *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 100–108, 2020.

[115] Zifeng Wang, Zheng Zhan, Yifan Gong, Geng Yuan, Wei Niu, Tong Jian, Bin Ren, Stratis Ioannidis, Yanzhi Wang, and Jennifer Dy. Sparcl: Sparse continual learning on the edge. *Advances in Neural Information Processing Systems*, 35:20366–20380, 2022.

[116] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer, 2022.

[117] Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian triplet loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[118] Fei Wen, Mian Qin, Paul Gratz, and Narasimha Reddy. An fpga-based hybrid memory emulation system. In *2021 31st International Conference on Field-Programmable Logic and Applications (FPL)*, pages 190–196, 2021.

[119] Fei Wen, Mian Qin, Paul Gratz, and Narasimha Reddy. Openmem: Hardware/software cooperative management for mobile memory system. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 109–114, 2021.

[120] Fei Wen, Mian Qin, Paul V. Gratz, and A. L. Narasimha Reddy. Hardware memory management for future mobile hybrid memory systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 39(11):3627–3637, 2020.

[121] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *European Conference on Computer Vision*, 2016.

[122] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-Area Image Geolocalization with Aerial Reference Imagery. In *IEEE International Conference on Computer Vision*, 2015.

[123] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic Networks: Deep Translation and Rotation Equivariance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7168–7177, 2016.

[124] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Fastformer: Additive attention can be all you need. *arXiv preprint arXiv:2108.09084*, 2021.

[125] Jiancan Wu, Xiang Wang, Xingyu Gao, Jiawei Chen, Hongcheng Fu, Tianyu Qiu, and Xiangnan He. On the effectiveness of sampled softmax loss for item recommendation, 2022.

[126] Jun Wu, Xuesong Ye, and Yanyuet Man. Bottrinet: A unified and efficient embedding for social bots detection via metric learning. *arXiv preprint arXiv:2303.03144*, 2023.

[127] Jun Wu, Xuesong Ye, Chengjie Mou, and Weinan Dai. Fineehr: Refine clinical note representations to improve mortality prediction. *arXiv preprint arXiv:2304.11794*, 2023.

[128] Yuqing Xie, Taesik Na, Xiao Xiao, Saurav Manchanda, Young Rao, Zhihong Xu, Guanghua Shu, Esther Vasiete, Tejaswi Tenneti, and Haixun Wang. An embedding-based grocery search model at instacart. *arXiv preprint arXiv:2209.05555*, 2022.

[129] Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Wang, Taibai Xu, and Ed H. Chi. Mixed negative sampling for learning two-tower neural networks in recommendations. 2020.

[130] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[131] Aron Yu and Kristen Grauman. Thinking outside the pool: Active training image creation for relative attributes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[132] Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L. Berg, and Ning Zhang. Commercemm: Large-scale commerce multimodal representation learning with omni retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’22, page 4433–4442, New York, NY, USA, 2022. Association for Computing Machinery.

[133] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[134] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. Von Luxburg,

S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[135] Amir Roshan Zamir and Mubarak Shah. Accurate Image Localization Based on Google Maps Street View. In *European Conference on Computer Vision*, 2010.

[136] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting Ground-Level Scene Layout from Aerial Imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[137] Jianjin Zhang, Zheng Liu, Weihao Han, Shitao Xiao, Ruicheng Zheng, Yingxia Shao, Hao Sun, Hanqing Zhu, Premkumar Srinivasan, Denvy Deng, Qi Zhang, and Xing Xie. Uni-retriever: Towards learning the unified embedding based retriever in bing sponsored search, 2022.

[138] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[139] Hongke Zhao, Qi Liu, Yong Ge, Ruoyan Kong, and Enhong Chen. Group preference aggregation: A nash equilibrium approach. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 679–688, 2016.

[140] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[141] Yi Zhu, Xueqing Deng, and Shawn D. Newsam. Fine-Grained Land Use Classification at the City Scale Using Ground-Level Images. In *IEEE Transactions on Multimedia*, 2018.