Relieving Triplet Ambiguity: Consensus Network for Language-Guided Image Retrieval

Xu Zhang
CCAI, Zhejiang University
Hangzhou, China
xu.zhang@zju.edu.cn

Zhedong Zheng
Sea-NExT Joint Lab, National University of Singapore
Singapore, Singapore
zdzheng12@gmail.com

Xiaohan Wang
CCAI, Zhejiang University
Hangzhou, China
wxh1996111@gmail.com

Yi Yang
CCAI, Zhejiang University
Hangzhou, China
yangyics@zju.edu.cn

ABSTRACT

Language-guided image retrieval enables users to search for images and interact with the retrieval system more naturally and expressively by using a reference image and a relative caption as a query. Most existing studies mainly focus on designing image-text composition architecture to extract discriminative visual-linguistic relations. Despite great success, we identify an inherent problem that obstructs the extraction of discriminative features and considerably compromises model training: **triplet ambiguity**. This problem stems from the annotation process wherein annotators view only one triplet at a time. As a result, they often describe simple attributes, such as color, while neglecting fine-grained details like location and style. This leads to multiple false-negative candidates matching the same modification text. We propose a novel Consensus Network (Css-Net) that self-adaptively learns from noisy triplets to minimize the negative effects of triplet ambiguity. Inspired by the psychological finding that groups perform better than individuals, Css-Net comprises 1) a consensus module featuring four distinct compositors that generate diverse fused image-text embeddings and 2) a Kullback-Leibler divergence loss, which fosters learning among the compositors, enabling them to reduce biases learned from noisy triplets and reach a consensus. The decisions from four compositors are weighted during evaluation to further achieve consensus. Comprehensive experiments on three datasets demonstrate that Css-Net can alleviate triplet ambiguity, achieving competitive performance on benchmarks, such as +2.77% R@10 and +6.67% R@50 on FashionIQ.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; **Image representations**; **Natural language processing**.

KEYWORDS

Representation Learning, Multi-modal Retrieval, Image Retrieval with Text Feedback, Triplet Ambiguity.

Figure 1: Illustration of the language-guided image retrieval system. Using a reference image and an associated descriptive sentence, the system endeavors to accurately retrieve the intended target image from candidate images for user convenience.

1 INTRODUCTION

Image retrieval is a fundamental task in computer vision and proves to be valuable in many applications, such as product search [21, 22, 49], internet search [43] and fashion retrieval [36, 40]. Prevalent image retrieval approaches include image-to-image retrieval [12–14, 50, 55] and text-to-image retrieval [7, 19, 29, 57, 66, 68], which endeavor to locate the image of interest using a single image or descriptive texts as a query. Despite significant progress in image retrieval, users often lack a precise search target in advance but instead seek categories, such as shoes or clothing. Therefore, an interactive system is highly desirable to assist users to reconsider their intentions, as depicted in Fig. 1. Hence, Language-guided image retrieval, which aims to search the target image of interest given the composed query consisting of a reference image and the relative caption describing the desired modification, has recently attracted great attention [6, 30, 35, 53, 58].

Recent studies addressing the task of language-guided image retrieval primarily concentrate on extracting discriminative representations from image-text-image triplets. For example, TIRG [53], VAL [6], and CoSMo [35] propose different ways to modify the

Xu Zhang, Zhedong Zheng, Xiaohan Wang, and Yi Yang

Figure 2: Illustration of the triplet ambiguity problem. Triplet ambiguity denotes multiple false-negative samples in the dataset. It is due to the annotator usually seeing one triplet with true match (✅) at a time, while neglecting other candidates (❓) in the whole dataset. Triplet ambiguity largely compromise the traditional metric learning on pushing away other negatives.

visual features of the reference image conditioned on the relative caption. TIRG uses a simple gating and residual module, VAL devises a visual-linguistic attention learning framework, and CoSMo introduces the content and style modulators. Additionally, CLVC-Net [58] and CLIP4cir [1] devise more intricate multi-modal fusion modules to accentuate the modifications of the reference image. CLVC-Net uses local-wise and global-wise composition modules, while CLIP4cir finetunes the CLIP [47] text encoder and trains a combiner network to fuse the visual and textual features.

Despite the significant success, these works fail to address an inherent problem of the language-guided image retrieval task: the ambiguity of the training data triplets, *i.e.*, **triplet ambiguity**. Triplet ambiguity originates from the annotation process in which annotators, focusing on single data triplet, frequently describe simple properties such as color, while neglecting more fine-grained details, such as location and style. Consequently, many noisy triplets exist where candidate images meet the requirement of the composed query but are not annotated as the desired ground-truth target image, especially when the relative caption is brief. Fig. 2 shows examples that apart from the true match marked with ✅, the other two candidate images marked with ❓ could also serve as the target image of the composed query. The noisy triplets lead to multiple false-negative candidates capable of fulfilling the same modification text, compromising the representation learning of the model. It is because the metric learning objective in this task aims to push away these false-negative samples from the composed query. We empirically verify that triplet ambiguity does exist in the language-guided image retrieval task in Sec. 4.2. Specifically, we compare the batch-based classification (mostly used in previous works) with a global-wise classification. We find that such a global-wise classification significantly degrades the performance, validating our assumption on triplet ambiguity.

To address the triplet ambiguity problem, we propose a straightforward and effective Consensus Network (Css-Net) for language-guided image retrieval, as illustrated in Fig. 3(a). The key idea underpinning our method to alleviate the triplet ambiguity is "two heads are better than one" in short. To be more specific, an individual often errs due to inherent biases, but groups are less susceptible to making

similar mistakes, thereby circumventing sub-optimal solutions. This is known as the psychological finding [27] that groups perform better than individuals on the memory task. Consequently, our goal is to (1) develop a consensus module composed of compositors possessing diverse knowledge to jointly make decisions during evaluation and (2) encourage learning among the compositors to minimize their biases learned on noisy triplets by employing an additional Kullback Leibler divergence loss (KL loss) [33].

To ensure that the compositors possess distinct knowledge, we differentiate them in two ways: • Motivated by the finding [37, 42] that the image features of high-resolution are semantically weak, while the image features of low-resolution are semantically strong, we first employ two image-text compositors at different depths of the same image encoder, (*i.e.*, block3 and block4 of the ResNet [25]). The former focuses more on detailed change like "has a purple star pattern", while the latter emphasizes more overall change such as "is modern and fashional". • Unlike the image-text compositor that uses relative caption to describe **what should change** on the reference image, we devise the text-image compositor to capture the textual cues based on text-to-image retrieval, where the reference image implies **what should preserve**. Specifically, we denote the reference image feature as f_r, the text feature as f_s, and the composed feature as $\hat{g}$. The image-text compositors primarily devised by previous works [6, 35, 53, 58] are in the residual form of $\hat{g}_{IT} = f_r + comp(f_r, f_s)$, where $comp$ represents a function to fuse f_r and f_s. The proposed text-image compositors are in the form of $\hat{g}_{TI} = f_s + comp(f_s, f_r)$ for capturing the textual cues of the query. We incorporate two symmetric text-image compositors at the same depths of the image encoder as image-text compositors. These four compositors share the same image and text encoders but exhibit distinct feature representations based on their respective knowledge. They collaboratively make decisions during evaluation to mitigate the individual biases learned on noisy triplets, which enhances the language-guided image retrieval performance.

To further reduce the negative impact of triplet ambiguity, we impose an additional KL loss between two image-text compositors. The KL loss enables two compositors to learn from each other and reach a consensus. This **soft label** combined with the respective **knowledge** from two compositors is more effective than the supervision from one-hot labels, as it helps each compositor to mitigate its own bias learned on noisy triplets and thus prevents the overfitting to the annotated target image. To demonstrate that KL loss provides additional information for compositors, we employ an intuitive label-smoothing approach as the soft label. However, we find that the uniform distribution of soft labels without knowledge does not address the triplet ambiguity due to the high false positive rate problem. In comparison, the KL loss bridges two image-text compositors more flexibly and feasibly. The experimental results show that Css-Net has achieved competitive performance on three benchmarks, empirically validating the effectiveness of our method.

In summary, our contributions are as follows:

• We identify an inherent problem in the language-guided image retrieval task and further verify the phenomenon through the preliminary experiments. We observe that the triplet ambiguity leads to sub-optimal model learning (*see Fig. 4*).

• To address triplet ambiguity, we introduce a Consensus Network (Css-Net) featuring a consensus module with four unique compositors for joint inference (*see Table 3*). Moreover, we employ KL loss to facilitate learning among compositors and reduce their biases learned on noisy triplets, making Css-Net more robust to triplet ambiguity. *See results in Table 2.*

• Extensive experiments show that the proposed method minimizes the negative impacts of noisy triplets. On three prevalent public benchmarks, we observe that Css-Net significantly surpasses the current state-of-the-art competitive methods, *e.g.*, with +2.77% Recall@10 on Shoes, and +6.67% Recall@50 on FashionIQ (*see Table 4, 5, and 6*).

2 RELATED WORK

2.1 Cross-modal Image Retrieval

Cross-modal image retrieval is a fundamental task in computer vision that has attracted wide attention from researchers. The most popular patterns of image retrieval are image-to-image matching [9, 12, 38, 51, 55, 59] and text-to-image matching [34, 64, 68], which allow users to search for images of interest with a similar image or some descriptive texts as queries. Although these paradigms have made great progress, they do not provide enough convenience for users to express their search intention. Therefore, more forms of image retrieval with flexible queries such as sketch-based image retrieval [11, 20, 52, 54] have emerged. In this work, we focus on the language-guided image retrieval task which involves a composed query of a reference image and a corresponding caption. To tackle this task, recent works [5, 6, 17, 35, 53, 58, 61, 63, 65] aim to devise a composition architecture to capture the visual-linguistic relation. For example, TIRG [53] uses a simple gating and residual module, VAL [6] devises a visual-linguistic attention learning framework, and CoSMo [35] introduces the content and style modulators. Besides, CLVC-Net [58] devises local-wise and global-wise composition modules, resembling model ensemble. Unlike the methods described above, our Css-Net does not rely on complicated composition modules for learning. Instead, our Css-Net mainly focuses on alleviating the triplet ambiguity problem that leads to a sub-optimal solution for a single compositor. To address this problem, Css-Net trains a consensus module to infer during evaluation and leverage KL loss to reduce individual bias during training.

2.2 Attention Mechanism

The attention mechanism is widely used in language and vision tasks in machine learning to capture the long-range dependencies and the relations between features. This mechanism is also inspired by a psychological finding [8] that humans observe and pay attention to specific parts as needed. In the language-guided image retrieval task, many works use the attention mechanism to design the image-text compositor. For example, VAL [6] employs self-attention by concatenating the text feature to each location of the image features. CoSMo [35] adopts the disentangled multi-modal non-local block to stabilize the training procedure of the content modulator. Besides, CLVC-Net [58] proposes a complex cross-attention between the feature of each word in the sentence and each spatial location of the image feature. In our work, we focus on utilizing several compositors with different knowledge. Without loss of generalizability, we deploy the

widely-used CoSMo [35] as our image-text compositor, which takes the feature map of the reference image and the pooled text feature (sentence-level feature) as input. Moreover, we propose a unique text-image compositor to fully utilize the attention mechanism to capture the relation of the average pooled reference image feature and the word-level text feature, which is orthogonal with existing attention-based models and could further improve the performance.

2.3 Co-training

Co-training is a semi-supervised learning technique that exploits two classifiers to acquire complementary information on two views of the data [3]. It has been extensively utilized in various research fields such as image recognition [46], semantic segmentation [44] and domain adaptation [41, 48, 67]. For instance, in domain adaptation, these co-training works explicitly maximize the discrepancies of the classifiers by utilizing extra losses such as adversarial loss [48] or weight discrepancy loss [41]. In contrast, our work adopts a co-training paradigm that leverages four compositors with different knowledge to jointly make decisions for the language-guided image retrieval task. We do not introduce extra loss to explicitly maximize the discrepancy of the four compositors, as they inherently possess various knowledge due to their different designs. For example, the two image-text compositors focus on the detailed and overall changes to the reference images based on the perspective of finding "what should change" in the reference image, and two text-image compositors are in view of the text-to-image retrieval with the reference image implying "what should preserve". Instead, we explicitly encourage the consensus between compositors and leverage the consensus to rectify the single prediction, which is aligned with this work [18] exploring the consistent and complementary correlations of multi-modal data. Refer to Sec. 3.2 for more details.

3 METHOD

This section describes the Consensus Network in detail. Sec. 3.1 introduces the overall framework of the network. Sec. 3.2 elaborates on the consensus module consisting of four distinct compositors with diverse knowledge, and the triplet ambiguity resolution. Sec. 3.3 discusses Css-Net and some recent and relevant works.

3.1 Overview of Consensus Network

As illustrated in Fig. 3 (a), the Consensus Network consists of three components: the image encoder, the text encoder, and the consensus module. The image encoder, F_{img}, extracts mid-level and high-level representations of the input images as:

$$f_r^m, f_r^h = F_{img}(I_r), \tag{1}$$

where I_r is the reference image, and $f_r^m, f_r^h \in \mathbb{R}^{C_{in} \times (H \times W)}$ refer to the mid-level and high-level image feature, respectively (*i.e.*, output from block3 and block4 of the ResNet [25]). Note that, since f_r^m and f_r^h are not used in the same compositor, the symbols with the superscript m in the subsequent formulas correspond to f_r^m rather than f_r^h, and vice versa. $C_{in} \times (H \times W)$ represents the shape of the feature maps. For brevity, we do not distinguish between different shapes of the image feature maps. The text encoder, denoted as F_{text},
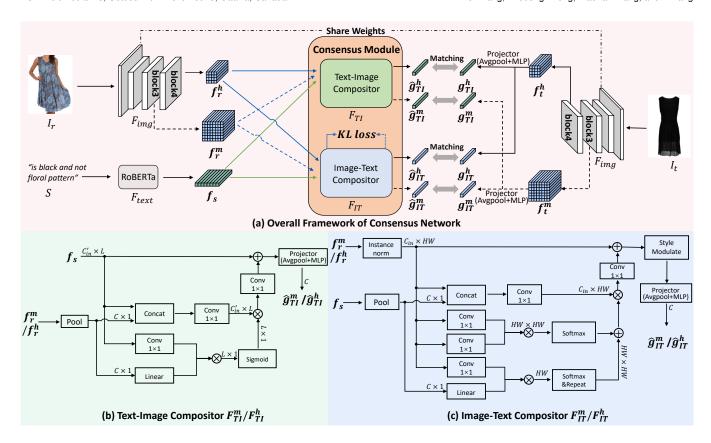
Figure 3: The overview of the Consensus Network. Given a reference image and a relative caption, we first extract the mid-level image feature f_r^m and high-level image feature f_r^h with the image encoder F_{img}, and the text representation f_s with the text encoder F_{text}. Then, we fuse the text representation with either the mid-level or high-level image feature using compositors. The solid blue lines represent mid-level image features, while dotted blue lines indicate high-level image features and solid green lines denote text features. The text-image compositor F_{TI} has the residual form of $f_s + F(f_s, f_r)$, which takes the word-level text representation f_s and the average pooled reference image feature $pool(f_r)$ as input. The image-text compositor F_{IT} has the residual form of $f_r + F(f_r, f_s)$ taking the intermediate feature map of the reference image f_r and the average pooled sentence-level text feature $pool(f_s)$ as input. Each compositor generates its own composed feature. We use a simple attention-based multi-modal non-local block for the text-image compositor and employ CoSMo [35] for the image-text compositor. Finally, we match the composed features with the target image feature to train the model. The projector block consists of an averaging pooling layer (Avgpool) and a multilayer perceptron (MLP).

extracts the features of the relative caption as follows:

$$f_s = F_{text}(S), \quad (2)$$

where S denotes the relative caption, $f_s \in \mathbb{R}^{C'_{in} \times L}$ refers to the word-level representation, and L is the number of words in the caption.

After extracting the image and text features, the consensus module transforms the reference image features with the corresponding text features into the composed features. It consists of four distinct compositors possessing different knowledge. These compositors at different depths of the image encoder can be grouped into two types. Specifically, given the reference image feature f_r and the text feature f_s, the composed query $\hat{g}$ can be obtained by either an image-text compositor or a text-image compositor. The image-text compositor has the residual form of $\hat{g}_{IT} = f_r + comp(f_r, f_s)$, which focuses on "what should change" to the reference image, while the text-image

compositor has the residual form of $\hat{g}_{TI} = f_s + comp(f_s, f_r)$ and emphasizes "what should preserve" based on the text-to-image retrieval. Here, *comp* represents a function to fuse f_r and f_s. Considering both the performance and computational efficiency, the two text-image compositors F_{TI}^m, F_{TI}^h, shown in Fig. 3 (b), take the word-level representation f_s along with the average pooled reference image features $pool(f_r^m), pool(f_r^h)$ as input, respectively, which are given by:

$$\begin{cases} \hat{g}_{TI}^m = F_{TI}^m(f_s, pool(f_r^m)) \\ \hat{g}_{TI}^h = F_{TI}^h(f_s, pool(f_r^h)), \end{cases} \quad (3)$$

where $\hat{g}_{TI}^m, \hat{g}_{TI}^h \in \mathbb{R}^C$ are the composed features from text-image compositors. Similarly, the image-text compositors F_{IT}^m, F_{IT}^h, shown in Fig. 3 (c) take the intermediate image feature maps, f_r^m, f_r^h along with the pooled sentence-level text representation $pool(f_s)$ as input,

which are given by:

$$\begin{cases} \hat{g}_{IT}^m = F_{IT}^m(f_r^m, pool(f_s)) \\ \hat{g}_{IT}^h = F_{IT}^h(f_r^h, pool(f_s)), \end{cases} \tag{4}$$

where $\hat{g}_{IT}^m$, $\hat{g}_{IT}^h \in \mathbb{R}^C$ are the composed features from image-text compositors. The target image features f_t^m, f_t^h are obtained from the same image encoder F_{img} as the reference image features. Then the projector blocks (composed of an average pooling layer and a multilayer perceptron (MLP)) are employed to acquire the target features: $g_{TI}^m, g_{TI}^h, g_{IT}^m$, and g_{IT}^h. The four compositors are trained by reducing the distance between the composed features and their corresponding projected target features within the embedding space.

In the next section, we will explain how these diverse compositors with different knowledge in the consensus module learn to relieve the triplet ambiguity problem.

3.2 Consensus Module

To address the triplet ambiguity, we propose the consensus module that consists of four distinct compositors with different knowledge. These compositors are trained to generate the composed query $\hat{g}$ that is close to the corresponding target image feature g in the feature space. At the evaluation stage, compositors independently compute the similarity between each composed query and all candidate target images in the gallery and collaboratively rank the whole gallery by aggregating the given similarities with different weights. We initially discuss the design of ensuring each compositor acquires distinct knowledge, followed by elucidating how compositors learn from each other to reduce their biases learned on noisy triplets.

3.2.1 Pyramid Training for Image-Text Compositor. We develop a pyramid training paradigm for image-text compositors, which is inspired by the finding [37, 42] that the image features of high-resolution are semantically weak, while the image features of low-resolution are semantically strong. Through exploring the different spatial information of the reference image, the two image-text compositors F_{IT}^m and F_{IT}^h independently learn unique knowledge by leveraging the batch-based classification loss, as given by:

$$\mathcal{L}_{IT}^m = -\log \frac{\exp(\hat{g}_{IT}^m \cdot g_{IT,+}^m)}{\sum_{j=1}^{B} \exp(\hat{g}_{IT}^m \cdot g_{IT,j}^m)} \tag{5}$$

and

$$\mathcal{L}_{IT}^h = -\log \frac{\exp(\hat{g}_{IT}^h \cdot g_{IT,+}^h)}{\sum_{j=1}^{B} \exp(\hat{g}_{IT}^h \cdot g_{IT,j}^h)}, \tag{6}$$

where $\hat{g}_{IT}^m$ and $\hat{g}_{IT}^h$ are mid-level and high-level composed features from two image-text compositors (Eq. 4). $g_{IT,+}^m$ and $g_{IT,+}^h$ are corresponding target features obtained from different projectors. The independent batch-based classification loss makes each image-text compositor learn from the interactions between text and different spatial information of the image, which enables these compositors to hold unique knowledge.

3.2.2 Auxiliary knowledge from Text-Image Compositor. The text-image compositor is a brand-new framework for generating the composed feature from the reference image and text, which is seldom referred to in previous works. It offers additional knowledge

due to its distinct design from the image-text compositor. The text-image compositor mainly focuses on the text-to-image retrieval with the reference image implying "what should preserve", while the image-text compositor finds "what should change" in the reference image. We use two symmetric text-image compositors at the same depths of the image encoder to capture different knowledge. The compositors reuse the features from the image encoder F_{img} and the text encoder F_{text} with minimal cost. We also apply a batch-based classification loss for each compositor F_{TI}^m and F_{TI}^h, as follows:

$$\mathcal{L}_{TI}^m = -\log \frac{\exp(\hat{g}_{TI}^m \cdot g_{TI,+}^m)}{\sum_{j=1}^{B} \exp(\hat{g}_{TI}^m \cdot g_{TI,j}^m)} \tag{7}$$

and

$$\mathcal{L}_{TI}^h = -\log \frac{\exp(\hat{g}_{TI}^h \cdot g_{TI,+}^h)}{\sum_{j=1}^{B} \exp(\hat{g}_{TI}^h \cdot g_{TI,j}^h)}, \tag{8}$$

where $\hat{g}_{TI}^m$ and $\hat{g}_{TI}^h$ are composed features from two text-image compositors (Eq. 3), respectively. $g_{TI,+}^m$ and $g_{TI,+}^h$ are corresponding target features obtained from different projector blocks.

3.2.3 Collaborative Consensus Learning. The triplet ambiguity problem causes the compositors to learn from noisy triplets and introduces biases. To mitigate this problem, we use the Kullback Leibler divergence loss (KL loss) for two image-text compositors. The KL loss enables the compositors to learn collaboratively from each other, reducing biases and reaching a consensus. This approach balances the preservation of distinct knowledge and the achievement of consensus. By enhancing cooperation and knowledge sharing, our method is more robust to the triplet ambiguity problem. Specifically, we denote the resulting posterior probability of F_{IT}^m as p^m and that of F_{IT}^h as p^h. We set a target probability p^w as the weighted sum of both p^m and p^h, which is given by:

$$p^w = \lambda_1 \cdot p^m + \lambda_2 \cdot p^h, \tag{9}$$

where λ_1 and λ_2 are two weight coefficients, and thus the KL loss is formulated as:

$$\mathcal{L}_{KL} = D_{KL}(p^m||p^w) + D_{KL}(p^h||p^w), \tag{10}$$

where D_{KL} is the KL divergence distance. The batch-based classification loss and KL loss play complementary roles in our approach. The application of KL loss minimizes individual biases of compositors with distinct knowledge. It is not essential to incorporate additional KL loss for the two text-image compositors, given their similarities in input. Specifically, both text-image compositors receive pooled reference image features with identical dimensions and share the same text representations. Consequently, the primary function of these text-image compositors is to act as auxiliary decision-makers during joint inference, addressing the triplet ambiguity issue. The final loss for training is the sum of the above loss functions:

$$\mathcal{L} = \mathcal{L}_{IT}^m + \mathcal{L}_{IT}^h + \mathcal{L}_{TI}^m + \mathcal{L}_{TI}^h + \mathcal{L}_{KL} \tag{11}$$

3.2.4 Joint Inference. We train four distinct compositors to independently learn different knowledge from the data triplets and enable the knowledge transfer to reduce biases learned on noisy triplets. At the evaluation step, we involve each compositor in decision-making to further minimize individual bias. Specifically, we use each compositor to independently generate composed features and

measure the similarity between any composed feature and target feature. The resulting similarity matrices are denoted as P_{IT}^m, P_{IT}^h, P_{TI}^m, $P_{TI}^h \in \mathbb{R}^{n_1 \times n_2}$, where n_1 and n_2 are the number of queries and target images in the gallery. The final similarity matrix for ranking the gallery is the weighted sum of four similarity matrices from distinct compositors:

$$P = \alpha_1 \cdot P_{IT}^m + \alpha_2 \cdot P_{IT}^h + \alpha_3 \cdot P_{TI}^m + \alpha_4 \cdot P_{TI}^h, \qquad (12)$$

where $\alpha_1 \ldots \alpha_4$ are weight coefficients. Note that a common practice that concatenates multiple composed features as one query is a special case that all α are equal to 1.

3.3 Discussions

VAL [6] and CLVC-Net [58] are most relevant to our Css-Net. Although VAL employs hierarchical matching strategies, our Css-Net diverges fundamentally in three respects: 1) Facilitating knowledge sharing among compositors at various depths for consensus, as opposed to independent compositors of VAL; 2) Omitting the low-level compositor to enhance performance and efficiency; 3) Implementing a weighted sum during evaluation, enabling adjustable influence of compositors. CLVC-Net incorporates global-wise and local-wise learning through two distinct models, akin to a model ensemble. Conversely, compositors in Css-Net utilize the same encoders but acquire unique knowledge from data triplets, employing a co-training strategy that renders it both effective and efficient. In conclusion, Css-Net represents a novel language-guided image retrieval approach that leverages compositors to learn diverse knowledge from noisy triplets, shares knowledge across compositors to minimize biases, and distinguishes itself from VAL and CLVC-Net.

4 EXPERIMENTS

This section consists of four parts. Sec. 4.1 describes the experimental setup. Sec. 4.2 presents the empirical evidence of the triplet ambiguity problem. Sec. 4.3 conducts some diagnostic experiments for our Consensus Network. Sec. 4.4 evaluates the performance of our method and compares it with the state-of-the-art methods. More qualitative results are provided in Supplementary Material.

4.1 Experimental Setup

4.1.1 Datasets. We evaluate our method on three large-scale language-guided image retrieval datasets: Shoes [2], FashionIQ [60], and Fashion200k [53].

The Shoes dataset [2] is originally crawled from *like.com* for attribute discovery. It is then annotated in the form of a triplet for dialog-based interactive retrieval. We follow VAL [6] to use $10,000$ training samples and $4,658$ evaluation samples.

The FashionIQ dataset [60] is a language-based interactive fashion retrieval dataset with $77,684$ images across three categories: Dresses, Tops&Tees, and Shirts. It includes $18,000$ triplets from $46,609$ training images, each containing a reference image, a target image, and two descriptive natural language captions. The evaluation procedure follows VAL [6] and CoSMo [35] for a fair comparison.

The Fashion200k dataset [23] contains over $200k$ fashion images from various websites and is for attribute-based product retrieval. With descriptive attributes for each product, $172k$ images are used for training and $33,480$ test queries for evaluation, following VAL

and CoSMo methods. Attributes generate relative descriptions using an online-processing pattern. As shown in Figure 2, we observe that Fahsion200k also meets the triplet ambiguity problem.

4.1.2 Implementation Details. We modify CoSMo [35] as our baseline by replacing LSTM [16] with RoBERTa [39] as the text encoder. ResNet-50 [25] serves as the image encoder for Shoes and FashionIQ datasets, while ResNet-18 [25] is used for Fashion200k. Image encoders are pretrained on ImageNet. Embedding space dimension C is 512, and the output sizes of image feature maps $C_{in} \times (H \times W)$ for ResNet50 are $1024 \times (14 \times 14)$ and $2048 \times (7 \times 7)$. Text feature shape is $C'_{in} \times L$, with C'_{in} being 768 and L is the sentence length. During training, we set $\lambda_1 = 10$ and $\lambda_2 = 1$, while evaluation uses $\alpha_1 \ldots \alpha_4 = 1, 0.5, 0.5, 0.5$. We adopt the standard evaluation metric in retrieval, *i.e.*, Recall@K, denoted as R@K for short.

We use Adam [32] as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. On Shoes and FashionIQ datasets, the batch size is 30 and the base learning rate of the text encoder and other modules are $2e - 6$ and $2e - 5$, respectively. On Fashion200k, the batch size is 126 and the base learning rate for the text encoder and other modules are $2e - 6$ and $2e - 4$, respectively. We adopt the warm-up scheme for the first 5 epochs. The learning rate decays at epoch 35 and epoch 45 by a factor of 10, and the total number of epochs is 50.

4.2 Triplet Ambiguity Verification

4.2.1 Global-wise v.s. Batch-based Optimization. To verify the negative impacts from the noisy triplets as shown in Fig. 2, we quantitatively compare between global-wise and batch-based optimization objectives. In particular, we mainly adopt ● Batch-based Classification (BBC): Limited negatives in the current batch are involved, and ● Global-wise Classification (GWC): Mining more negative samples in the training set for comparison.

If the data triplet does **NOT** have ambiguity, the global-wise classification has the potential to be a comparable even better choice since it uses more negative samples in the training set and potentially learns a better metric, which is consistent with some findings in metric learning [26, 50, 56] and self-supervised learning [4, 24]. Specifically, Consider a composed query q and a set of features/prototypes of the candidate target images $\{k_0, k_1, ...\}$, there is one true match denoted k_+ in the candidates. The two losses are given by:

$$\mathcal{L}_{BBC} = -\log \frac{\exp(q \cdot k_+))}{\sum_{i=1}^{B} \exp(q \cdot k_i))} \qquad (13)$$

and

$$\mathcal{L}_{GWC} = -\log \frac{\exp(q \cdot k_+))}{\sum_{i=1}^{N} \exp(q \cdot k_i))}, \qquad (14)$$

where B is the batch size, and N is the number of IDs (classes) in the training set. The only difference between them is that $\mathcal{L}_{GWC}$ involves more negative counterparts, which results in high false negative rates if the triplet ambiguity does exist. We conduct experiments on the Shoes dataset [2] using two losses, respectively, under the same settings of CoSMo [35]. We observe that batch-based methods outperform global-wise methods by a large margin, as shown in Fig. 4. The experimental results confirm our triplet ambiguity assumption: the training data contains many noisy triplets (*i.e.*, false negative samples) as Sec. 1 discusses, which makes learning on noisy triplets challenging. Although batch-based classification suffers less
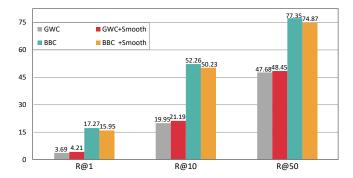
Figure 4: Comparison between the batch-based classification and the global-wise classification on the Shoes dataset. Batch-based classification discriminates different objects within the batch, while global-wise classification distinguishes all categories, introducing more ambiguous negatives. Obviously, the global-wise classification significantly degrades the performance since more false negative samples from triplet ambiguity are involved.

from triplet ambiguity, the single compositor still faces some noisy negative triplets in the batch and produces a sub-optimal solution.

4.2.2 Label Smoothing. We first briefly illustrate one intuitive way we consider to alleviate the triplet ambiguity problem: label smoothing. The motivation is that there are many false negative samples due to the triplet ambiguity and label smoothing could alleviate the overfitting to the annotated true match. In label smoothing, the label $y = [y_1, \ldots y_n]$ is not a hard one-hot label rather than a soft one-hot label, which is given by:

$$y_i = \begin{cases} 1 \ (if \ i = c) \\ 0 \ (if \ i \neq c) \end{cases} \Longrightarrow y_i = \begin{cases} 1 - \varepsilon \ (if \ i = c) \\ \frac{\varepsilon}{B-1} \ (if \ i \neq c), \end{cases} \quad (15)$$

where y_i is the label for class i, c is the corresponding class of the query, B is the batch size, and ε is a hyperparameter for label smoothing and is set to be 0.1. We use label smoothing for both the batch-based classification and the global-wise classification, and perform the experiments on the Shoes dataset, which are presented in Fig. 4. The experimental results indicate that label smoothing deteriorates the performance of batch-based classification but enhances the performance of global-wise classification. This is because ● global-wise classification is severely affected by triplet ambiguity since there are always false negative samples during learning, while batch-based classification is affected only when noisy negative triplets are in the batch; ● Label smoothing could alleviate the triplet ambiguity but also introduce another problem that many true negative target samples are assigned weights to learn, which impairs the model training for batch-based classification. The experimental results also verify the effectiveness of KL loss used in this work.

4.3 Diagnostic Experiments

4.3.1 Pyramid Training. In Sec. 3.2.1, we present the design of the pyramid training, which exploits the image features from the mid-level and high-level blocks of the image encoder. In this section, we verify its effectiveness by comparing it with different designs. Table 1 reports the experimental results. Our base model is $F_{IT}^m + F_{IT}^h$ used

Table 1: Comparison of various pyramid training methods on the Shoes dataset, which are trained and evaluated independently.

Method	Shoes		
	R@1	R@10	R@50
F_{IT}^h	17.27	52.26	77.35
$F_{IT}^l + F_{IT}^h$	18.24	52.14	78.12
$F_{IT}^l + F_{IT}^m + F_{IT}^h$	18.81	54.21	79.55
$F_{IT}^m + F_{IT}^h$	19.10	54.69	79.63

Table 2: Efficacy of model designs. The experiments are conducted on the Shoes dataset under the same setting.

$\mathcal{L}_{IT}^m$	$\mathcal{L}_{TI}^h + L_{TI}^m$	$\mathcal{L}_{KL}$	Shoes		
Eq. 6	Eq. 7&8	Eq. 10	R@1	R@10	R@50
Baseline: only $\mathcal{L}_{IT}^h$ (Eq. 5)			17.27	52.26	77.35
✓			19.10(+1.83)	54.69(+2.43)	79.63(+2.28)
✓	✓		19.47(+2.20)	54.63(+2.37)	80.46(+3.11)
✓	✓	✓	20.13(+2.86)	56.81(+4.55)	81.32(+3.97)

Table 3: Effect of joint inference. We train the whole model on the Shoes dataset and separately evaluate each component.

Inference Method	Shoes		
	R@1	R@10	R@50
F_{IT}^m	15.72	51.17	78.89
F_{IT}^h	18.35	55.15	80.52
F_{TI}^m	17.06	53.35	78.92
F_{TI}^h	16.58	52.17	77.77
Joint Inference (Eq. 12)	20.13	56.81	81.32

in Css-Net, which applies pyramid training on mid-level and high-level features. We conduct experiments on two variants for pyramid training: 1) $F_{IT}^l + F_{IT}^h$, which uses the image features from block2 and block4 of the ResNet, and 2) $F_{IT}^l + F_{IT}^m + F_{IT}^h$, which uses three image-text compositors at three blocks to generate the composed query. Both variants perform worse than Css-Net, e.g., −2.55% and −0.48% on the R@10 evaluation metric. However, they both surpass F_{IT}^h using only one image-text compositor at block4. These results indicate that 1) the low-level image feature is too semantically weak for pyramid training, and 2) groups perform better than individuals.

4.3.2 Efficacy of Model Designs. Table 2 shows the effectiveness of our core idea, which uses four different compositors with KL loss to address the triplet ambiguity problem. We make three observations from the table. First, employing image-text compositors at other layers of the image encoder (*i.e.*, $\mathcal{L}_{IT}^m$ in Eq. 6) can reduce the triplet ambiguity problem and improve the performance significantly (77.35% → 79.63% at R@50 metric). This indicates that two image-text compositors can benefit from the interactions between the relative caption and different spatial information of the reference image. Second, adding a new compositor framework, text-image compositor, to this task (*i.e.*, $\mathcal{L}_{TI}^m + \mathcal{L}_{TI}^h$ in Eq. 7&8) can further

Table 4: Quantitative results of language-guided image retrieval on the FashionIQ dataset. The best results are in bold. They symbol * marks an updated version by the same authors. The symbol † indicates that this method deploys model ensemble.

Method	Image Encoder	Dress		Shirt		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
MRN [31]	ResNet-152	12.32	32.18	15.88	34.33	18.11	36.33	15.44	34.28
FiLM [45]	ResNet-50	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04
TIRG [53]	ResNet-17	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39
VAL [6]	ResNet-50	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04
DCNet [30]	ResNet-50	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89
CoSMo* [35]	ResNet-50	26.45	52.43	26.94	52.99	31.95	62.09	28.45	55.84
CLVC-Net† [58]	ResNet-50×2	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41
ARTEMIS [10]	ResNet-50	27.16	52.40	21.78	54.83	29.20	43.64	26.05	50.29
CLIP4Cir [1]	ResNet-50	31.73	56.02	35.77	57.02	36.46	62.77	34.65	58.60
Baseline	ResNet-50	30.95	56.98	31.48	59.98	36.97	67.31	33.13	61.42
Css-Net	ResNet-50	**33.65**	**63.16**	**35.96**	**61.96**	**42.65**	**70.70**	**37.42**	**65.27**

Table 5: Quantitative results on the Shoes dataset. The best results are in bold. The symbol † denotes model ensemble method.

Method	Shoes		
	R@1	R@10	R@50
MRN [31]	11.74	41.70	67.01
FiLM [45]	10.19	38.89	68.30
TIRG [53]	12.6	45.45	69.39
VAL [6]	16.49	49.12	73.53
CoSMo [35]	16.72	48.36	75.64
DCNet [30]	-	53.82	79.33
CLVC-Net† [58]	17.64	54.39	79.47
ARTEMIS [10]	18.72	53.11	79.31
Baseline	17.27	52.26	77.35
Css-Net	**20.13**	**56.81**	**81.32**

Table 6: Quantitative results on Fashion200k dataset. The best results are in bold. The symbol † denotes model ensemble method.

Method	Fashion200k		
	R@1	R@10	R@50
MRN [31]	13.4	40.0	61.9
FiLM [45]	12.9	39.5	61.9
TIRG [53]	14.1	42.5	63.8
VAL [6]	21.2	49	68.8
DCNet [30]	-	46.9	67.6
CoSMo [35]	23.3	50.4	69.3
CLVC-Net† [58]	22.6	**53.0**	**72.2**
ARTEMIS [10]	21.5	51.1	70.5
Baseline	20.9	47.7	67.8
Css-Net	22.2	50.5	69.7
Css-Net†	**23.4**	52.0	72.0

improve the performance (79.63% → 80.46% at R@50 metric). This demonstrates the advantage of the novel text-image compositors. Third, applying an extra KL loss for the posterior probability from two image-text compositors ($\mathcal{L}_{KL}$ in Eq. 10) can enhance the performance notably (80.46% → 81.32% at R@50 metric). This suggests that the KL loss enables two image-text compositors to share and learn from their respective knowledge, thus minimizing the biases.

4.3.3 Effect of Joint Inference. At the evaluation stage, Css-Net makes compositors jointly make the decision as introduced in Sec. 3.2.4. Table 3 shows the experimental results. It is observed that joint inference surpasses every single compositor and verifies our motivation that groups perform better than individuals and could be used to reduce their own biases mainly caused by triplet ambiguity.

4.4 The Effectiveness of Our Method

We present the experimental results in Table 4, Table 5, and Table 6. We could make two observations: **(1) We adopt a competitive baseline with few modifications.** As mentioned in Sec. 4.1, we adopt the CoSMo as our baseline and replace the LSTM with a more robust text encoder: RoBERTa, and observe consistent improvement. For example, on the FashionIQ dataset, our baseline improves CoSMo by 4.68% R@10 on average, and surpasses CoSMo by 3.90% R@10 on

the Shoes dataset. We infer that RoBERTa is more robust than LSTM [28] to more accurately capture the textual information. However, our baseline is slightly lower than the reported results of CoSMo on Fashion200k, as the authors do not provide sufficient implementation details for reproducing. This also limits comparing our method with CQBIR [62], whose baseline uses faster RCNN [15] as a different image encoder. Nevertheless, our method is more effective than CQBIR on FashionIQ and Shoes, where the triplet ambiguity problem is more serious. **(2) The proposed Css-Net could further improve and advances the state of the art on such a strong baseline, verifying the effectiveness of Css-Net.** For example, Table 4 shows Css-Net improves retrieval accuracy on all FashionIQ subsets. Compared to the baseline, it gains +2.70% R@10 on Dress, +4.48% R@10 on Shirt, and +5.68% R@10 on TopTee. Compared to previous works, our method brings overall improvements (e.g., +2.77% R@10 and +6.67% R@50 on average by CLIP4Cir). The improvements are significant and empirically validate the effectiveness of Css-Net for handling the triplet ambiguity problem. Besides in Table 5, Css-Net surpasses the state-of-the-art (CLVC-Net) on the Shoes dataset, achieving improvements of +2.49% R@1 and +2.42% R@10, which further demonstrates that Css-Net is robust across different

datasets. Table 6 presents Fashion200k results. Although our baseline is below the reported results of CosMo because of insufficient implementation details for reproduction, Css-Net brings a considerable improvement (*e.g.*, +2.8% R@10 over the baseline) and is still competitive with many state-of-the-art works especially when applying the model ensemble (*e.g.*, +4.3% R@10 by the baseline).

5 CONCLUSION

We present a Consensus Network (Css-Net) for language-guided image retrieval. Css-Net aims to relieve the inherent triplet ambiguity problem, which arises when the dataset contains multiple false-negative candidates that match the same query text. This problem stems from annotators overlooking fine-grained details of the images and describing only simple properties. The resulting noisy triplets significantly compromise the metric learning objective. To alleviate this problem, Css-Net employs a consensus module with four diverse compositors that possess different knowledge and can learn mutually during training and infer collaboratively when evaluation. Specifically, Css-Net adopts a pyramid training paradigm and auxiliary text-image compositor design that endow each compositor with unique knowledge. Css-Net also utilizes a KL loss that facilitates the learning among the compositors and reduces their biases learned on noisy triplets. Our experiments show that Css-Net is a competitive method on three benchmarks, demonstrating its effectiveness and robustness. Moreover, Css-Net is orthogonal and complementary to most existing methods, and can further enhance their performance. As future work, we plan to extend our method to real-world applications that involve learning from noisy triplets.

REFERENCES

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4959–4968.
[2] Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*. Springer, 663–676.
[3] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. 92–100.
[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
[5] Yanbei Chen and Loris Bazzani. 2020. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 136–152.
[6] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3001–3011.
[7] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8415–8424.
[8] Maurizio Corbetta and Gordon L Shulman. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience* 3, 3 (2002), 201–215.
[9] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-modality person re-identification with generative adversarial training.. In *IJCAI*, Vol. 1. 2.
[10] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *International Conference on Learning Representations*.
[11] Cheng Deng, Xinxun Xu, Hao Wang, Muli Yang, and Dacheng Tao. 2020. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. *IEEE Transactions on Image Processing* 29 (2020), 8892–8902.
[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4690–4699.
[13] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. 2019. Spherereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation* 60 (2019), 51–58.
[14] Fangxiang Feng, Tianrui Niu, Ruifan Li, Xiaojie Wang, and Huixing Jiang. 2020. Learning Visual Features from Product Title for Image Retrieval *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 4723–4727.
[15] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.
[16] Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.
[17] Chunbin Gu, Jiajun Bu, Zhen Zhang, Zhi Yu, Dongfang Ma, and Wei Wang. 2021. Image Search with Text Feedback by Deep Hierarchical Attention Mutual Information Maximization. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 4600–4609.
[18] Weili Guan, Haokun Wen, Xuemeng Song, Chung-Hsing Yeh, Xiaojun Chang, and Liqiang Nie. 2021. Multimodal Compatibility Modeling via Exploring the Consistent and Complementary Correlations. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 2299–2307.
[19] Ricardo Guerrero, Hai X Pham, and Vladimir Pavlovic. 2021. Cross-modal retrieval and synthesis (x-mrs): Closing the modality gap in shared subspace learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3192–3201.
[20] Longteng Guo, Jing Liu, Yuhang Wang, Zhonghua Luo, Wei Wen, and Hanqing Lu. 2017. Sketch-Based Image Retrieval Using Generative Adversarial Networks. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 1267–1268.
[21] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yinglong Wang, Jun Ma, and Mohan Kankanhalli. 2019. Attentive long short-term preference modeling for personalized product search. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–27.
[22] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. 2018. Multi-Modal Preference Modeling for Product Search *(MM '18)*. Association for Computing Machinery, New York, NY, USA, 1865–1873.
[23] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*. 1463–1471.
[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.
[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[26] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).
[27] Verlin B Hinsz. 1990. Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social psychology* 59, 4 (1990), 705.
[28] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[29] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. 2017. Cross-Domain Image Retrieval with Attention Modeling *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 1654–1662.
[30] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1771–1779.
[31] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. *Advances in neural information processing systems* 29 (2016).
[32] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[33] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.
[34] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 201–216.
[35] Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 802–812.
[36] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable Multimodal Retrieval for Fashion Products *(MM '18)*. Association

for Computing Machinery, New York, NY, USA, 1571–1579.

[37] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[38] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. 2022. Learning Memory-Augmented Unidirectional Metrics for Cross-Modality Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19366–19375.

[39] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[40] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1096–1104.

[41] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2507–2516.

[42] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9826–9836.

[43] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*. 3456–3465.

[44] Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. 2020. Deep co-training for semi-supervised image segmentation. *Pattern Recognition* 107 (2020), 107269.

[45] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[46] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*. 135–152.

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

[48] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3723–3732.

[49] Rishab Sharma and Anirudha Vishvakarma. 2019. Retrieving similar e-commerce images using deep learning. *arXiv preprint arXiv:1901.03546* (2019).

[50] Hao Sheng, Yanwei Zheng, Wei Ke, Dongxiao Yu, Xiuzhen Cheng, Weifeng Lyu, and Zhang Xiong. 2020. Mining hard samples globally and efficiently for person reidentification. *IEEE Internet of Things Journal* 7, 10 (2020), 9611–9622.

[51] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6398–6407.

[52] Jialin Tian, Xing Xu, Zheng Wang, Fumin Shen, and Xin Liu. 2021. Relationship-Preserving Knowledge Distillation for Zero-Shot Sketch Based Image Retrieval *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 5473–5481.

[53] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6439–6448.

[54] Hao Wang, Cheng Deng, Tongliang Liu, and Dacheng Tao. 2021. Transferable Coupled Network for Zero-Shot Sketch-Based Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1.

[55] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5265–5274.

[56] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. 2020. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6388–6397.

[57] Zheng Wang, Zhenwei Gao, Xing Xu, Yadan Luo, Yang Yang, and Heng Tao Shen. 2022. Point to Rectangle Matching for Image Text Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 4977–4986.

[58] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive linguistic-visual composition network for image retrieval. In

Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1369–1378.

[59] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-Infrared Cross-Modality Person Re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 5390–5399.

[60] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11307–11317.

[61] Yuchen Yang, Min Wang, Wengang Zhou, and Houqiang Li. 2021. Cross-modal joint prediction and alignment for composed query image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3303–3311.

[62] Feifei Zhang, Ming Yan, Ji Zhang, and Changsheng Xu. 2022. Comprehensive Relationship Reasoning for Composed Query Based Image Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 4655–4664.

[63] Gangjian Zhang, Shikui Wei, Huaxin Pang, and Yao Zhao. 2021. Heterogeneous feature fusion and cross-modal alignment for composed image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5353–5362.

[64] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3536–3545.

[65] Yida Zhao, Yuqing Song, and Qin Jin. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1012–1021.

[66] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10394–10403.

[67] Zhedong Zheng and Yi Yang. 2019. Unsupervised scene adaptation with memory regularization in vivo. *IJCAI* (2019).

[68] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.

Figure 5: Top-10 retrieval results on three datasets. The queries consist of a reference image and a relative caption that describes the desired modification. The blue/green boxes refer to the reference image and the true match(es) in the database.

APPENDIX

A FURTHER QUALITATIVE ANALYSIS

Figure 5 shows the top-10 retrieval results on three datasets. We make three key observations from these results: (1) Css-Net can capture the information of the reference image and the relative caption for both coarse-grained and fine-grained queries. For example, the first query of Shoes and the third query of FashionIQ retrieve the correct matches easily, and the first and second queries of FashionIQ also find the correct matches. (2) The model sometimes fails to retrieve the correct matches due to the triplet ambiguity problem, *e.g.*, the first query of Fashion200K retrieves some negative samples but are highly related to the query. (3) Css-Net is less sensitive to some detailed information such as location. For example, the third query in Shoes retrieves a shoe that is visually similar but has a wrong paid location, because the dataset has few similar training samples. Improving the model's sensitivity to the detailed information is a direction for our future work.