Collaborative Group: Composed Image Retrieval via Consensus Learning from Noisy Annotations

Xu Zhang[a], Zhedong Zheng[b,*], Linchao Zhu[a], Yi Yang[a]

[a] *College of Computer Science and Technology, Zhejiang University, Hangzhou, 310058, China*
[b] *Faculty of Science and Technology, and Institute of Collaborative Innovation, University of Macau, Macau, 999078, China*

Abstract

Composed image retrieval extends content-based image retrieval systems by enabling users to search using reference images and captions that describe their intention. Despite great progress in developing image-text compositors to extract discriminative visual-linguistic features, we identify a hitherto overlooked issue, triplet ambiguity, which impedes robust feature extraction. Triplet ambiguity refers to a type of semantic ambiguity that arises between the reference image, the relative caption, and the target image. It is mainly due to the limited representation of the annotated text, resulting in many noisy triplets where multiple visually dissimilar candidate images can be matched to an identical reference pair (i.e., a reference image + a relative caption). To address this challenge, we propose the Consensus Network (Css-Net), inspired by the psychological concept that groups outperform individuals. Css-Net comprises two core components: (1) a consensus module with four diverse compositors, each generating distinct image-text embeddings, fostering complementary feature extraction and mitigating dependence on any single, potentially biased compositor; (2) a Kullback-Leibler divergence loss that encourages learning of inter-compositor interactions to promote consensual outputs. During evaluation, the decisions of the four compositors are combined through a weighting scheme, enhancing overall agreement. On benchmark datasets, particularly FashionIQ, Css-Net demonstrates marked improvements. Notably, it achieves significant recall gains, with a 2.77% in-

*Corresponding author.
Email address: zhedongzheng@um.edu.mo (Zhedong Zheng)

crease in R@10 and 6.67% boost in R@50, underscoring its competitiveness in addressing the fundamental limitations of existing methods.

Keywords: Noisy Annotation, Data Ambiguity, Compositional Image Retrieval, Image Retrieval with Text Feedback, Multi-modal Retrieval

1. Introduction

Image retrieval plays a pivotal role in computer vision and proves to be valuable in many applications, such as product search (Guo et al., 2019; Sharma and Vishvakarma, 2019; Guo et al., 2018), internet search (Noh et al., 2017) and fashion retrieval (Liu et al., 2016; Liao et al., 2018). Prevalent image retrieval approaches include image-to-image retrieval (Deng et al., 2019; Fan et al., 2019; Sheng et al., 2020; Hafner et al., 2022) and text-to-image retrieval (Zhen et al., 2019; Zheng et al., 2020; Guerrero et al., 2021; Wang et al., 2022), which endeavor to locate the image of interest using a single image or descriptive texts as a query. Despite significant progress, 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, Composed image retrieval, which aims to search the image of interest given the composed query consisting of a reference image and a relative caption describing the modification, has attracted great attention (Vo et al., 2019; Chen et al., 2020; Lee et al., 2021; Kim et al., 2021; Wen et al., 2021).

Recent studies addressing the task of composed image retrieval primarily concentrate on extracting discriminative representations from image-text-image triplets. For example, TIRG (Vo et al., 2019), VAL (Chen et al., 2020), and CoSMo (Lee et al., 2021) propose different ways to modify the 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 (Wen et al., 2021) and CLIP4cir (Baldrati et al., 2022) devise more intricate multi-modal fusion modules to accentuate the modifications of the reference image. CLVC-Net uses local-wise and global-wise compositors, while CLIP4cir finetunes the CLIP (Radford et al., 2021) text encoder and trains a combiner network to fuse features.

Despite the significant success, these works fail to address an inherent problem of the composed image retrieval task: the ambiguity of the training

Figure 1: Schematic illustration of the composed image retrieval system. Through using a reference image and a relative caption, the system endeavors to precisely retrieve the intended target image from all candidate images.

data triplets, *i.e.*, **triplet ambiguity**. Triplet ambiguity originates from the annotation process where annotators focus on single data triplet, and frequently describe simple properties such as color and size, 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. Similar annotation ambiguity is also observed in pair-wise data (Wray et al., 2021; Falcon et al., 2022) and remains challenging. As shown in Fig. 2, existing methods treat composed image retrieval as an instance-level retrieval, that is, given a reference pair (comprising a reference image and a relative caption), only the annotated target image is considered as the correct image to retrieve. In fact, due to the limitation of the text description, many candidate images within the dataset are semantically similar to the point of being identical, but are treated as the negative counterparts, thus producing many noisy triplets. These noisy triplets compromises the representation learning of the single compositor, since the metric learning objective in this task aims to push away these false-negative samples from the composed query. Empirically, we

Figure 2: Illustration of the triplet ambiguity problem. Triplet ambiguity denotes multiple false-negative samples in the dataset as the annotator usually see one triplet with true match (✓) at a time, while neglecting other candidates (?).

verify the existence of triplet ambiguity in Sec. 4.2.

To relieve the triplet ambiguity problem, we propose a straightforward and effective Consensus Network (Css-Net) for composed 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 the biases caused by noisy triplets, but groups are less susceptible to making similar mistakes, thereby circumventing sub-optimal solutions. This is known as the psychological finding (Hinsz, 1990) that groups perform better than individuals on the memory task. Consequently, we aim to (1) develop a consensus module (group) composed of compositors possessing diverse knowledge to jointly make decisions during evaluation and (2) encourage learning among different compositors to minimize their biases learned on noisy triplets by employing an additional Kullback Leibler divergence loss (KL loss) (Kullback and Leibler, 1951).

Css-Net ensures that the compositors possess distinct knowledge in two ways: • Motivated by the finding (Lin et al., 2017; Miech et al., 2021) that the image features of high-resolution are semantically weak, while the image features of low-resolution are semantically strong, we employ two image-text compositors at different depths of the same image encoder, (*i.e.*, block3 and block4 of the ResNet (He et al., 2016)). 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 to 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 to preserve** for the referenece image. See details in Sec. 3.1. To minimize the negative impact of triplet ambiguity during training, we impose a KL loss between two image-text compositors. The KL loss promotes two compositors to learn from each other and reach a consensus, which is similar to supervision from peers in a group, as it helps each compositor to reduce its own bias and thus avoids overfitting to the annotated target image.

In summary, our contributions are as follows:

• We have identified an inherent issue within the context of composed image retrieval, namely triplet ambiguity, which we subsequently confirm through initial experimental investigations (*see Fig. 2 and 4*). This problem, stemming from the inherent noisiness of the annotation process, results in suboptimal model learning, as it compromises the extraction of discriminative features that integrate visual and linguistic information.

• To relieve triplet ambiguity, we introduce the Consensus Network (Css-Net) featuring a consensus module with four distinct compositors for collaborative training (*see Table 5*) and joint inference (*see Table 6*).

• 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 1, 2, and 3*).

2. Related Work

Cross-modal Image Retrieval. Cross-modal image retrieval has attracted wide attention from researchers. The most popular patterns of image retrieval are image-to-image matching (Zheng et al., 2017; Deng et al., 2019; Sun et al., 2020; Wu et al., 2017; Dai et al., 2018; Liu et al., 2022; Qu et al., 2024) and text-to-image matching (Liu et al., 2019; Zhang et al., 2020; Liu et al., 2022; Zhang et al., 2022; Li et al., 2024). 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 (Deng et al., 2020; Wang et al., 2021; Li et al., 2022; Liang et al., 2024) have emerged. In this work, the composed image retrieval task involves a composed query of a reference image and a relative caption. To tackle this task, recent works (Vo et al., 2019; Chen et al., 2020; Yang et al., 2021; Zhang et al., 2021; Lee et al., 2021; Wen et al., 2021; Gu et al., 2021; Zhao et al., 2022; Han et al., 2023) devise diverse composition architectures to capture the visual-linguistic relation. Unlike the methods described above, our Css-Net does not propose complicated compositors. Instead, our work mainly focuses on reducing single compositor biases to alleviate the identified triplet ambiguity problem.

Attention Mechanism. The attention mechanism is widely used in language and vision tasks in machine learning to capture the relations between features. This mechanism is also inspired by a psychological finding (Corbetta and Shulman, 2002) that humans observe and pay attention to specific parts as needed. In the composed image retrieval task, many works use the attention mechanism to design the image-text compositor. For example, VAL (Chen et al., 2020) employs self-attention to capture the image-text relations by concatenating the text feature to the image feature. CoSMo (Lee et al., 2021) adopts the disentangled multi-modal non-local block to stabilize the training procedure for learning better representations. Besides, CLVC-Net (Wen et al., 2021) proposes a cross attention between each word in the sentence and each spatial location of the image feature to recognize details. In our work, the main idea is not to design a new attention-based compositor but to utilize several compositors to form as a consensus module. Without loss of generality, we deploy the widely-used CoSMo as the image-text compositor. Moreover, we propose specific text-image compositors based on cross attention to better capture the relation between the reference image feature and the word-level text feature, which is orthogonal with existing attention-based models and could further improve the retrieval performance.

Co-training. Co-training is a semi-supervised learning technique that exploits two components to acquire complementary information on two views of the data (Blum and Mitchell, 1998). It has been extensively utilized in various research fields such as image recognition (Qiao et al., 2018), segmentation (Peng et al., 2020; Hui et al., 2023) and domain adaptation (Saito et al., 2018; Zheng and Yang, 2019; Luo et al., 2019). Our work adopts a co-training paradigm that leverages four compositors with different knowledge to jointly make decisions for the composed image retrieval task. The two image-text compositors focus on the detailed and overall changes to the reference images based on the perspective of finding "what to change" in the reference image, and the two text-image compositors are in view of the text-to-image retrieval with the reference image implying "what to preserve" for the relative caption. The compositors hold diverse knowledge from different views of the data. Thus, we explicitly encourage the consensus between compositors and leverage the consensus to rectify the single prediction.

3. Method

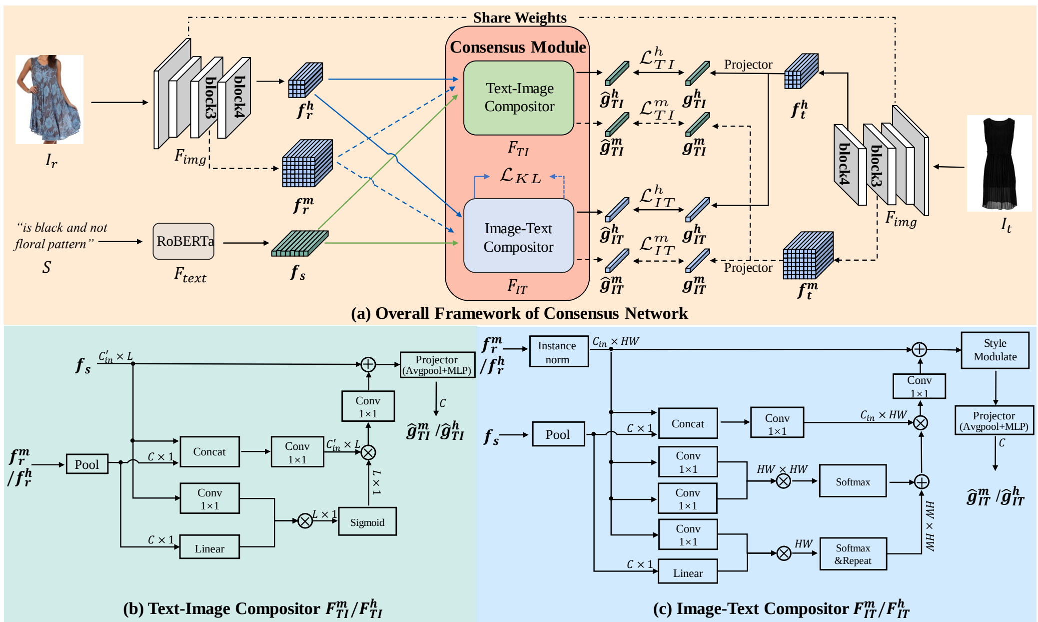

Figure 3: Schematic illustration of the Consensus Network. Given a reference image and a relative caption, the image encoder F_{img} extracts the mid-level image feature $\boldsymbol{f_r^m}$ and high-level image feature $\boldsymbol{f_r^h}$, and the text encoder F_{text} extracts the text feature $\boldsymbol{f_s}$. Then, compositors fuse the text feature with either the mid-level or high-level image feature. Each compositor generates distinct composed feature. Finally, we match the composed features with the corresponding target features and impose a KL loss between image-text compositors for training.

3.1. Overview of Consensus Network

As illustrated in Fig. 3 (a), the Consensus Network consists of three components: a image encoder, a text encoder, and a consensus module. The image encoder, F_{img}, extracts mid-level and high-level reference image features as $\boldsymbol{f_r^m}, \boldsymbol{f_r^h} = F_{img}(I_r)$, where I_r is the reference image, and

$\boldsymbol{f_r^m}, \boldsymbol{f_r^h} \in \mathbb{R}^{C_{in}\times(H\times W)}$ are mid-level and high-level image features, respectively (*i.e.*, output from block3 and block4 of the ResNet (He et al., 2016)). $C_{in} \times (H \times W)$ represents the shape of the feature maps. For brevity, we do not distinguish between different shapes of image features. The text encoder, denoted as F_{text}, extracts features of the relative caption as $\boldsymbol{f_s} = F_{text}(S)$, where S denotes the relative caption, $\boldsymbol{f_s} \in \mathbb{R}^{C'_{in}\times L}$ refers to the word-level representation, and L is the number of words of the relative 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 $\boldsymbol{f_r}$ and the text feature $\boldsymbol{f_s}$, the composed query $\boldsymbol{\hat{g}}$ can be obtained by either an image-text compositor or a text-image compositor. The image-text compositor has the residual form of $\boldsymbol{\hat{g}_{IT}} = \boldsymbol{f_r} + comp(\boldsymbol{f_r}, \boldsymbol{f_s})$ and mainly focuses on "what to change" for $\boldsymbol{f_r}$ conditioned on the relative caption, while the text-image compositor has the residual form of $\boldsymbol{\hat{g}_{TI}} = \boldsymbol{f_s} + comp(\boldsymbol{f_s}, \boldsymbol{f_r})$ and mainly emphasizes on "what to preserve" for $\boldsymbol{f_s}$ conditioned on the reference image. Here, *comp* represents a trained component to fuse $\boldsymbol{f_r}$ and $\boldsymbol{f_s}$ as the condition. Considering both the performance and computational efficiency, the text-image compositors F^m_{TI} and F^h_{TI}, shown in Fig. 3 (b), take the word-level representation $\boldsymbol{f_s}$ along with the average pooled reference image features $pool(\boldsymbol{f_r^m}), pool(\boldsymbol{f_r^h})$ as input, respectively:

$$\begin{cases} \boldsymbol{\hat{g}^m_{TI}} = F^m_{TI}(\boldsymbol{f_s}, pool(\boldsymbol{f_r^m})) \\ \boldsymbol{\hat{g}^h_{TI}} = F^h_{TI}(\boldsymbol{f_s}, pool(\boldsymbol{f_r^h})), \end{cases} \tag{1}$$

where $\boldsymbol{\hat{g}^m_{TI}}$, $\boldsymbol{\hat{g}^h_{TI}}$ are the composed features from text-image compositors. Similarly, the image-text compositors F^m_{IT}, F^h_{IT}, shown in Fig. 3 (c) take the intermediate image feature maps, $\boldsymbol{f_r^m}, \boldsymbol{f_r^h}$ along with the pooled sentence-level text representation $pool(\boldsymbol{f_s})$ as input, which are given by:

$$\begin{cases} \boldsymbol{\hat{g}^m_{IT}} = F^m_{IT}(\boldsymbol{f_r^m}, pool(\boldsymbol{f_s})) \\ \boldsymbol{\hat{g}^h_{IT}} = F^h_{IT}(\boldsymbol{f_r^h}, pool(\boldsymbol{f_s})), \end{cases} \tag{2}$$

where $\boldsymbol{\hat{g}^m_{IT}}$, $\boldsymbol{\hat{g}^h_{IT}}$ are the composed features from image-text compositors.

The target image features $\boldsymbol{f_t^m}, \boldsymbol{f_t^h}$ are obtained from the same image encoder F_{img} as the reference image features $\boldsymbol{f_r^m}, \boldsymbol{f_r^h}$. Then four independent projector blocks (composed of an average pooling layer and a MLP) are employed to acquire target features: $\boldsymbol{g^m_{TI}}$, $\boldsymbol{g^h_{TI}}$, $\boldsymbol{g^m_{IT}}$, and $\boldsymbol{g^h_{IT}}$. Finally, the four compositors are trained by pulling close the corresponding target while pushing away other negatives within the embedding space.

3.2. Consensus Module

To relieve the triplet ambiguity, we introduce the consensus module, which consists of four distinct compositors with different knowledge. These compositors have individual biases learned on noisy triplets, which are minimized at two stages. At the training stage, each compositor acquires information from different views of the data, and the KL loss enables them to learn from each other to minimize biases. At the evaluation stage, each compositor independently provides decisions and collaborates to rank the entire gallery by aggregating their decisions. We first discuss the design to ensure that each compositor acquires distinct knowledge, then explain how compositors learn from each other to reduce their biases learned on noisy triplets. The batch-based classification loss is as follows:

$$\mathcal{L}_{BBC} = -\log \frac{\exp(\hat{\boldsymbol{g}} \cdot \boldsymbol{g}_{+})}{\sum_{j=1}^{B} \exp(\hat{\boldsymbol{g}} \cdot \boldsymbol{g_j})}, \tag{3}$$

where $\hat{\boldsymbol{g}}$ is the composed feature from respective compositor, and $\boldsymbol{g_j}$ are candidates, among which the true match is $\boldsymbol{g}_{+}$.

Pyramid Training for Image-Text Compositor. We develop a pyramid training paradigm for image-text compositors, which is inspired by the finding (Lin et al., 2017; Miech et al., 2021) 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^m_{IT} and F^h_{IT} independently learn knowledge by leveraging the batch-based classification loss $\mathcal{L}^m_{IT}$ and $\mathcal{L}^h_{IT}$. The independent batch-based classification loss makes each image-text compositor learn from the interactions between relative caption and different spatial information of the reference image, which enables these compositors to hold distinct knowledge from each other.

Auxiliary knowledge from Text-Image Compositor. The text-image compositor is a fancy component for generating the composed feature from the input, which is seldom referred to in previous works. It offers additional knowledge due to its distinct design from the image-text compositor. As discussed in Sec. 3.1, the text-image compositor views the data from another perspective, mainly focusing on the text-to-image retrieval with the reference image implying "what to preserve" conditioned on the text information, while the image-text compositor finds "what to change" in the reference image. We use two symmetric text-image compositors at the same depths of the image encoder, leveraging the batch-based classification loss $\mathcal{L}_{TI}^m$ and $\mathcal{L}_{TI}^h$.

Collaborative Consensus Learning. The triplet ambiguity problem leads to noisy triplets and biases the model learning. To mitigate this problem, we use the Kullback Leibler divergence loss (KL loss) for two image-text compositors. The KL loss promotes the compositors to learn from each other, reducing biases and reaching a consensus. This approach balances the preservation of distinct knowledge and the attainment 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 $\boldsymbol{p^m}$ and that of F_{IT}^h as $\boldsymbol{p^h}$. We set a target probability $\boldsymbol{p^w}$ as the weighted sum of both $\boldsymbol{p^m}$ and $\boldsymbol{p^h}$, which is given by:

$$\boldsymbol{p^w} = \lambda_1 \cdot \boldsymbol{p^m} + \lambda_2 \cdot \boldsymbol{p^h}, \tag{4}$$

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

$$\mathcal{L}_{KL} = D_{KL}(\boldsymbol{p^m} || \boldsymbol{p^w}) + D_{KL}(\boldsymbol{p^h} || \boldsymbol{p^w}), \tag{5}$$

where D_{KL} is the KL divergence distance. The KL loss reduces the biases of the compositors during training, which works alongside the batch-based classification loss in our approach. The preliminary experiments show that it is not essential to incorporate extra KL loss for the two text-image compositors. See Sec. 5.3 for a detaileed explanation. The final loss for training is the sum of the above loss functions:

$$\mathcal{L} = \mathcal{L}_{IT}^m + \mathcal{L}_{IT}^h + \mathcal{L}_{TI}^m + \mathcal{L}_{TI}^h + \mathcal{L}_{KL}, \tag{6}$$

where L_{IT}^m, L_{IT}^h, L_{TI}^m, and L_{TI}^h are batch-based classification loss used for independently training each image-text/text-image compositor F_{IT}^m, F_{IT}^h, F_{TI}^m, and F_{TI}^h. The superscript m indicates the mid-level input feature, while h denotes the high-level feature. The subscript IT indicates the image-text compositor, while the subscript TI denotes the text-image compositor.

Joint Inference. The four distinct compositors 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:

$$P = \alpha_1 \cdot P_{IT}^m + \alpha_2 \cdot P_{IT}^h + \alpha_3 \cdot P_{TI}^m + \alpha_4 \cdot P_{TI}^h, \tag{7}$$

where $\alpha_1 \dots \alpha_4$ are weight coefficients to balance the decisions from four compositors. Note that a common practice that concatenates multiple composed features as one query is a special case where all αs are equal to 1.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate Css-Net on three composed image retrieval datasets, *i.e.*, Shoes (Berg et al., 2010), FashionIQ (Wu et al., 2021), and Fashion200k (Vo et al., 2019).

- The Shoes dataset (Berg et al., 2010) 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 (Chen et al., 2020) to use 10,000 training samples and 4,658 evaluation samples.

- The FashionIQ dataset (Wu et al., 2021) 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 (Chen et al., 2020) and CoSMo (Lee et al., 2021).

- The Fashion200k dataset (Han et al., 2017) 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. The relative descriptions are generated from attributes using an online-processing pattern.

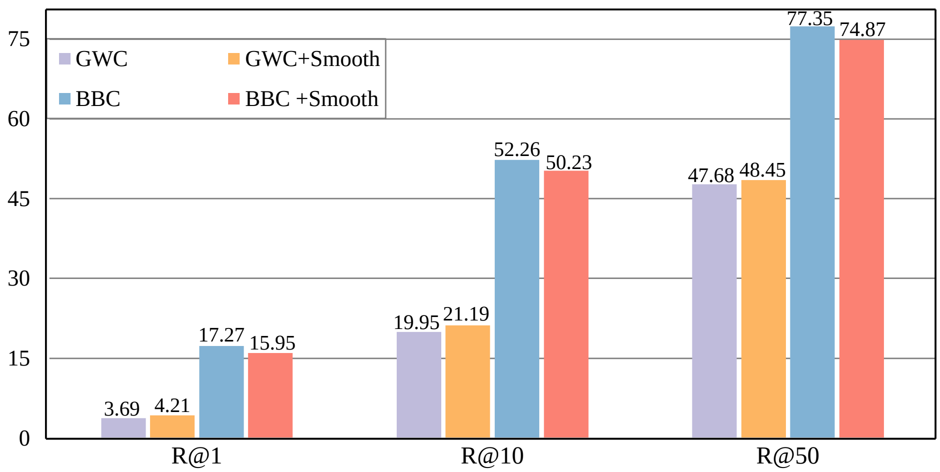

Figure 4: Comparison between the batch-based classification and the global-wise classification (GWC) on the Shoes dataset. GWC significantly degrades the performance since more false negative samples are involved due to triplet ambiguity.

4.2. Triplet Ambiguity Verification

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

If the data triplets do **NOT** have ambiguity, the global-wise classification has the potential to be comparable or even better since it uses more negative samples in the training set and potentially learns a better metric, which is consistent with many findings in metric learning (Hermans et al., 2017; Sheng et al., 2020; Wang et al., 2020) and self-supervised learning (Chen et al., 2020; He et al., 2020). Specifically, Given a query q and features/prototypes $\{k_0, k_1, ...\}$ of candidate target images, where the true match is denoted as k_+. Two losses are given by:

$$\mathcal{L}_{BBC} = -\log \frac{\exp(q \cdot k_+))}{\sum_{i=1}^{B} \exp(q \cdot k_i))} \tag{8}$$

and

$$\mathcal{L}_{GWC} = -\log \frac{\exp(q \cdot k_+))}{\sum_{i=1}^{N} \exp(q \cdot k_i))}, \tag{9}$$

Method	Dress		Shirt		Toptee		Average	
	R@10 ↑	R@50 ↑	R@10 ↑	R@50 ↑	R@10 ↑	R@50 ↑	R@10 ↑	R@50 ↑
MRN (Kim et al., 2016)	12.32	32.18	15.88	34.33	18.11	36.33	15.44	34.28
FiLM (Perez et al., 2018)	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04
TIRG (Vo et al., 2019)	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39
VAL (Chen et al., 2020)	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04
DCNet (Kim et al., 2021)	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89
CoSMo* (Lee et al., 2021)	26.45	52.43	26.94	52.99	31.95	62.09	28.45	55.84
CLVC-Net† (Wen et al., 2021)	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41
ARTEMIS (Delmas et al., 2022)	27.16	52.40	21.78	54.83	29.20	43.64	26.05	50.29
MUR (Chen et al., 2022)	30.60	57.46	31.54	58.29	37.37	68.41	33.17	61.39
CLIP4Cir (Baldrati et al., 2022)	31.73	56.02	35.77	57.02	36.46	62.77	34.65	58.60
Baseline	30.95	56.98	31.48	59.98	36.97	67.31	33.13	61.42
Css-Net	**33.65**	**63.16**	**35.96**	**61.96**	**42.65**	**70.70**	**37.42**	**65.27**

Table 1: Quantitative results on the FashionIQ dataset. The best results are in **bold**. The symbol * marks an updated results by the same authors. The symbol † indicates that this method deploys model ensemble (the same as below).

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 (Berg et al., 2010) using two losses, respectively, under the same settings of CoSMo (Lee et al., 2021). 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). Although batch-based classification suffers less from triplet ambiguity, the single compositor still faces some noisy negative triplets in the batch and produces a sub-optimal solution.

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

$$y_i = \begin{cases} 1 \ (if \ i = c) \\ 0 \ (if \ i \neq c) \end{cases} \Longrightarrow y_i = \begin{cases} 1 - \varepsilon \ (if \ i = c) \\ \frac{\varepsilon}{B-1} \ (if \ i \neq c), \end{cases} \tag{10}$$

where y_i is the label for class i, c is the corresponding class of the query, B

Method	Shoes		
	R@1 ↑	R@10 ↑	R@50 ↑
MRN (Kim et al., 2016)	11.74	41.70	67.01
FiLM (Perez et al., 2018)	10.19	38.89	68.30
TIRG (Vo et al., 2019)	12.60	45.45	69.39
VAL (Chen et al., 2020)	16.49	49.12	73.53
CoSMo (Lee et al., 2021)	16.72	48.36	75.64
DCNet (Kim et al., 2021)	-	53.82	79.33
CLVC-Net† (Wen et al., 2021)	17.64	54.39	79.47
MUR (Chen et al., 2022)	18.41	53.63	79.84
ARTEMIS (Delmas et al., 2022)	18.72	53.11	79.31
Baseline	17.27	52.26	77.35
Css-Net	**20.13**	**56.81**	**81.32**

Table 2: Quantitative results on the Shoes dataset. The best results are in **bold**. The symbol † indicates that this method deploys model ensemble. The proposed method has achieved competitive performances in all three metrics R@1, 10, 50.

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, 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 due to high false negative rate, while batch-based classification is affected only when noisy negative triplets are in the batch; • Label smoothing could relieve the triplet ambiguity but introduce another problem that many true negative target samples are assigned weights, which impairs the model training for batch-based classification. The experimental results also verify the effectiveness of KL loss as another form of soft label.

4.3. The Effectiveness of Our Method

We present the experimental results in Table 1, Table 2, and Table 3. 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

Method	Fashion200k		
	R@1 ↑	R@10 ↑	R@50 ↑
MRN (Kim et al., 2016)	13.4	40.0	61.9
FiLM (Perez et al., 2018)	12.9	39.5	61.9
TIRG (Vo et al., 2019)	14.1	42.5	63.8
VAL (Chen et al., 2020)	21.2	49	68.8
DCNet (Kim et al., 2021)	-	46.9	67.6
CoSMo (Lee et al., 2021)	23.3	50.4	69.3
CLVC-Net† (Wen et al., 2021)	22.6	**53.0**	**72.2**
ARTEMIS (Delmas et al., 2022)	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

Table 3: Quantitative results on the Fashion200k dataset. The best results are in **bold**. The symbol † indicates that this method deploys model ensemble. The proposed method has achieved competitive performances.

RoBERTa is more robust than LSTM (Hochreiter and Schmidhuber, 1997) to 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 (Zhang et al., 2022), whose baseline uses faster RCNN (Girshick, 2015) 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 1 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 2, 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 3 presents Fashion200k results. Although our baseline is

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 4: Comparison of various pyramid training methods on the Shoes dataset. These methods are trained and evaluated independently. F_{IT}^l, F_{IT}^m, and F_{IT}^h represent the low-level, mid-level, and high-level image-text compositor, respectively. The low-level compositor is useful, whereas the mid and high-level features show better performance.

$\mathcal{L}_{IT}^m$	$\mathcal{L}_{TI}^h + L_{TI}^m$	$\mathcal{L}_{KL}$	Shoes		
			R@1 ↑	R@10 ↑	R@50 ↑
Baseline: (only $\mathcal{L}_{IT}^h$)			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 5: Efficacy of model designs. L_{IT}^m, L_{IT}^h, L_{TI}^m, and L_{TI}^h are batch-based classification loss defined in Eqn. 3, and L_{KL} is the KL loss defined in Eqn. 5.

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 SOTA works especially when applying the model ensemble (*e.g.*, +4.3% R@10).

4.4. Diagnostic Experiments

Pyramid Training. In Sec. 3.2, 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. We verify its effectiveness by comparing it with different designs. Table 4 reports the experimental results. Our baseline is $F_{IT}^m + F_{IT}^h$ used in Css-Net. 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) $F_{IT}^l + F_{IT}^m + F_{IT}^h$, utilizing three image-text compositors at three depths. Both variants perform worse than Css-Net, e.g., −2.55% and −0.48% on the R@10 metric. However, they both surpass

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. 7)	**20.13**	**56.81**	**81.32**

Table 6: Effect of joint inference. We train Css-Net with four compositors on Shoes once and separately evaluate each compositor. Joint inference refers to using the weighting scheme (Eqn. 7) to combine decisions from all the compositors .

	Total time (s) ↓	Time per query (ms) ↓	Time per target (ms) ↓
Baseline (one)	168.2	50.2	56.4
Css-Net (four)	195.8	58.5	65.7

Table 7: Inference time cost for the baseline and Css-Net. Total time refers to the time taken to process all queries. Time per query indicates the average time spent on each query, while time per target represents the average time used to process each target in the gallery.

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 to provide image information, and 2) groups perform better than individuals.

Efficacy of Model Designs. Table 5 shows the effectiveness of our core idea, which uses four different compositors with KL loss to relieve 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.*, $\mathcal{L}_{IT}^m$) can mitigate 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 module, text-image compositor, to this task (*i.e.*, $\mathcal{L}_{TI}^m + \mathcal{L}_{TI}^h$) can further improve the performance (79.63% → 80.46% at R@50 metric). This demonstrates the advantage of auxiliary knowledge. Third, applying an extra KL loss for two image-text compositors ($\mathcal{L}_{KL}$) 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 their knowledge, thus minimizing the biases learned from noisy triplets.

Effect of Joint Inference At the evaluation stage, Css-Net allows compositors to jointly make the decision as introduced in Sec. 3.2. As shown in Table 6, joint inference surpasses single compositor and verifies our motivation that groups perform better than individuals and could be used to reduce their own prediction biases mainly caused by the triplet ambiguity problem.

Computational cost at inference Css-Net uses four compositors that share the same image and text encoders, thus adding minimal retrieval latency. The inference time is shown in Table 7, ranging from loading the model to displaying results. The experiments are conducted with GeForce RTX 2080 Ti, using $33,480$ queries and $29,789$ targets.

Implementation Details. We modify CoSMo (Lee et al., 2021) as our baseline by replacing LSTM (Graves, 2012) with RoBERTa (Liu et al., 2019) as the text encoder. ResNet-50 (He et al., 2016) serves as the image encoder for Shoes and FashionIQ datasets, while ResNet-18 (He et al., 2016) is used for Fashion200k. Embedding space dimension C is 512. 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 $\alpha_1 \dots \alpha_4 = 1, 0.5, 0.5, 0.5$. We adopt the standard evaluation metric in retrieval, *i.e.*, Recall@K, denoted as R@K for short. We use a random seed for each experiment and repeat it five times for the final results. we employ the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. On Shoes and FashionIQ, the batch size is set to be 32 and the base learning rates of the text encoder and other modules are $2e-6$ and $2e-5$, respectively. On Fashion200k, the batch size is set to be 128 and the base learning rates are $2e-6$ and $2e-4$, respectively. We adopt warm-up for the first 5 epochs, decay learning rate by 10 at epochs 35 and 45 during training. The total training epoch is 50.

5. Further Analysis and Discussion

5.1. Further Qualitative Analysis

Fig. 5 shows the top-10 retrieval results on three datasets: Shoes, Fashion200K, and FashionIQ. 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. These queries have clear and distinctive features that

Figure 5: Top-10 retrieval results on three datasets. The composed 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).

can be matched by Css-Net. (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 still 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 sensitivity of the model to the detailed information is a direction for our future work. We plan to explore more fine-grained features to enhance Css-Net in the future works.

5.2. Comparison with Most Relevant Works

We compare our Css-Net with VAL (Chen et al., 2020) and CLVC-Net (Wen et al., 2021), which are most relevant to our work.

(1) Our Css-Net differs from the hierarchical matching strategies in VAL:

- Our Css-Net facilitates knowledge sharing between compositors at various depths for consensus, instead of independent learning in VAL.
- Our Css-Net observes that the low-level compositor does not contribute to collaborative learning and omits it to enhance the recall performance and efficiency (*See Table 1*).
- Our Css-Net implements an adjustable weighted sum during evaluation, enabling individuals to make decisions as a group.

(2) Our Css-Net differs from the model ensemble design in CLVC-Net:

- Our Css-Net is more efficient since all compositors share the same encoder stem (Table 7), while model ensembling in CLVCNet employs several independent backbones.
- Our Css-Net encourages intra-modal and inter-modal knowledge sharing via collaborative learning between compositors, while model ensembling does not entail additional loss or learning among the models.
- Our Css-Net acknowledges that the compositors have different knowledge and thus assign adaptive weights, while model ensembling usually presumes that the models are independent and equally important.
- Our Css-Net enables single compositor to perform better could further benefits from model ensembling (See Table 6), while model ensembling does not improve single compositor prediction.

5.3. Discussion of Collaborative Learning

We apply a KL loss between text-image compositors in a preliminary experiment, but find that it is not as significant as the KL loss between image-text compositors. This is because the inputs for the text-image compositors are too similar, as shown in Fig. 6. Specifically, both text-image compositors receive a pooled reference image feature with identical dimensions and share the same text representations. Therefore, the main function of these text-image compositors is to act as auxiliary decision-makers during joint inference, addressing the triplet ambiguity issue. For simplicity and efficiency, we do not incorporate additional KL loss for the text-image compositors. However, we note that the text-image compositors still play an important role in our framework, as they provide complementary information to the image-text compositors and improve the retrieval performance.

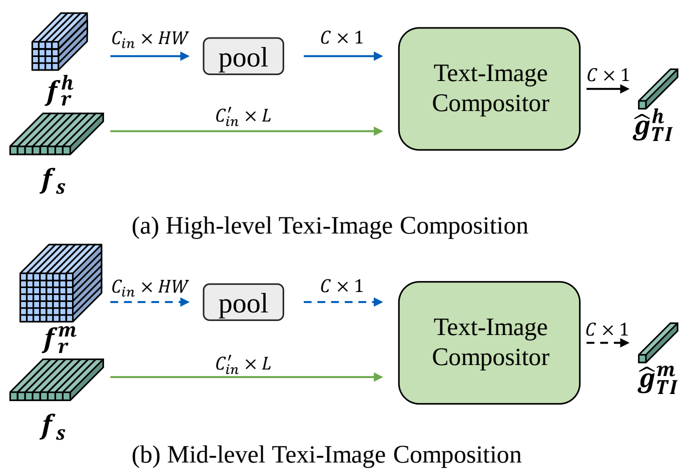

Figure 6: A brief illustration of two text-image compositors with the input shape. Please refer to Fig. 3 for the entire framework.

5.4. Analysis for hyperparameters

	R@1	R@10	R@50
Css-Net ($\alpha_{1-4} = 1, 1, 1, 1$)	20.04	56.44	80.87
Css-Net ($\alpha_{1-4} = 1, 0.5, 0.5, 0.5$)	20.13	56.81	81.32

Table 8: Ablation for hyperparameter α.

	R@1	R@10	R@50
Css-Net ($\lambda_{1-2} = 1, 1$)	19.95	56.64	80.55
Css-Net ($\lambda_{1-2} = 10, 1$)	20.13	56.81	81.32

Table 9: Ablation for Hyperparameter λ.

In this work, the hyperparameters αs and λs are not handpicked, as we empirically find that they are not sensitive and do not affect the model performance significantly. We set αs to be $1:0.5:0.5:0.5$ based on the observation that the high-level image-text compositor performs best among all compositors (Table 8) and we want this compositor to act like a leader in the group. Similarly, we use λs $= 1$ for all compositors, as we have some preliminary experiments that show similar results with this setting (Table 9). To demonstrate this, we add some experimental results on the Shoes dataset, which is another challenging benchmark for composed image retrieval. The results show that our Css-Net achieves competitive performance with different values of αs and λs, indicating that our model is robust and stable to the choice of hyperparameters.

5.5. Effect of More Annotation Noise

In this work, we aim to relieve the issue of noisy annotations, which can compromise the entire training process. Further, we artificially increased the noise intensity during training by manually manipulating relevant captions, such as random deletion, random swap, and random insertion proposed in a NLP work (Wei and Zou, 2019). To be more specific, we conducted an experiment on the Shoes dataset for both the baseline and Css-Net. For each relative caption, there is a 50% probability of adding one of three types of noise: Each word in the sentence has a 50% probability of being deleted; half of the words in the sentence are replaced with synonyms; and new words are inserted into half of the word intervals. The performance of the newly developed baseline and Css-Net are shown in Table 10.

Method	R@1	R@10	R@50
Baseline (*w* noise)	16.29	50.14	75.91
Css-Net (*w* noise)	19.07	55.69	78.98
Baseline (*w/o* noise)	17.27	52.26	77.35
Css-Net (*w/o* noise)	20.13	56.81	81.32

Table 10: Effect of annotation noise (w/o refers to without; w refers to with).

6. Conclusion

We present a Consensus Network (Css-Net) for composed 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. This problem stems from annotators describing only simple properties and frequently overlooking fine-grained details of the images. The resulting noisy triplets significantly compromise the metric learning objective and bias the single compositor. To this end, Css-Net employs a consensus module with four compositors that possess distinct knowledge. As a group, compositors learn mutually when training and infer collaboratively during evaluation, effectively minimizing the negative effects caused by the triplet ambiguity problem. Extensive experiments show that Css-Net has achieved competitive recall performance on three widely-used benchmarks, without substantially increasing the inference time.

References

Y. Guo, Z. Cheng, L. Nie, Y. Wang, J. Ma, M. Kankanhalli, Attentive long short-term preference modeling for personalized product search, ACM Transactions on Information Systems 37 (2019) 1–27.

R. Sharma, A. Vishvakarma, Retrieving similar e-commerce images using deep learning, arXiv:1901.03546 (2019).

Y. Guo, Z. Cheng, L. Nie, X.-S. Xu, M. Kankanhalli, Multi-modal preference modeling for product search, in: ACM Multimedia, 2018.

H. Noh, A. Araujo, J. Sim, T. Weyand, B. Han, Large-scale image retrieval with attentive deep local features, in: ICCV, 2017, pp. 3456–3465.

Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: CVPR, 2016, pp. 1096–1104.

L. Liao, X. He, B. Zhao, C.-W. Ngo, T.-S. Chua, Interpretable multimodal retrieval for fashion products, in: ACM Multimedia, 2018.

J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: CVPR, 2019, pp. 4690–4699.

X. Fan, W. Jiang, H. Luo, M. Fei, Sphereid: Deep hypersphere manifold embedding for person re-identification, Journal of Visual Communication and Image Representation 60 (2019) 51–58.

H. Sheng, Y. Zheng, W. Ke, D. Yu, X. Cheng, W. Lyu, Z. Xiong, Mining hard samples globally and efficiently for person reidentification, IEEE Internet of Things Journal 7 (2020) 9611–9622.

F. M. Hafner, A. Bhuyian, J. F. Kooij, E. Granger, Cross-modal distillation for rgb-depth person re-identification, Computer Vision and Image Understanding 216 (2022) 103352.

L. Zhen, P. Hu, X. Wang, D. Peng, Deep supervised cross-modal retrieval, in: CVPR, 2019, pp. 10394–10403.

Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Transactions on Multimedia Computing, Communications, and Applications 16 (2020) 1–23.

R. Guerrero, H. X. Pham, V. Pavlovic, Cross-modal retrieval and synthesis (x-mrs): Closing the modality gap in shared subspace learning, in: ACM Multimedia, 2021, pp. 3192–3201.

Z. Wang, Z. Gao, X. Xu, Y. Luo, Y. Yang, H. T. Shen, Point to rectangle matching for image text retrieval, in: ACM Multimedia, 2022.

N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, J. Hays, Composing text and image for image retrieval an empirical odyssey, in: CVPR, 2019, pp. 6439–6448.

Y. Chen, S. Gong, L. Bazzani, Image search with text feedback by visiolinguistic attention learning, in: CVPR, 2020, pp. 3001–3011.

S. Lee, D. Kim, B. Han, Cosmo: Content-style modulation for image retrieval with text feedback, in: CVPR, 2021, pp. 802–812.

J. Kim, Y. Yu, H. Kim, G. Kim, Dual compositional learning in interactive image retrieval, in: AAAI, volume 35, 2021, pp. 1771–1779.

H. Wen, X. Song, X. Yang, Y. Zhan, L. Nie, Comprehensive linguistic-visual composition network for image retrieval, in: SIGIR, 2021, pp. 1369–1378.

A. Baldrati, M. Bertini, T. Uricchio, A. Del Bimbo, Conditioned and composed image retrieval combining and partially fine-tuning clip-based features, in: CVPR, 2022, pp. 4959–4968.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021, pp. 8748–8763.

M. Wray, H. Doughty, D. Damen, On semantic similarity in video retrieval, in: CVPR, 2021, pp. 3650–3660.

A. Falcon, G. Serra, O. Lanz, A feature-space multimodal data augmentation technique for text-video retrieval, in: ACM Multimedia, 2022, pp. 4385–4394.

V. B. Hinsz, Cognitive and consensus processes in group recognition memory performance., Journal of Personality and Social psychology 59 (1990) 705.

S. Kullback, R. A. Leibler, On information and sufficiency, The annals of mathematical statistics 22 (1951) 79–86.

T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: CVPR, 2017, pp. 2117–2125.

A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, A. Zisserman, Thinking fast and slow: Efficient text-to-visual retrieval with transformers, in: CVPR, 2021, pp. 9826–9836.

K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, ACM transactions on multimedia computing, communications, and applications 14 (2017) 1–20.

Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle loss: A unified perspective of pair similarity optimization, in: CVPR, 2020, pp. 6398–6407.

A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, Rgb-infrared cross-modality person re-identification, in: ICCV, 2017, pp. 5390–5399.

P. Dai, R. Ji, H. Wang, Q. Wu, Y. Huang, Cross-modality person re-identification with generative adversarial training., in: IJCAI, volume 1, 2018, p. 2.

J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, W. Li, Learning memory-augmented unidirectional metrics for cross-modality person re-identification, in: CVPR, 2022, pp. 19366–19375.

X. Qu, L. Liu, L. Zhu, L. Nie, H. Zhang, Source-free style-diversity adversarial domain adaptation with privacy-preservation for person re-identification, Knowledge-Based Systems 283 (2024) 111150.

R. Liu, Y. Zhao, S. Wei, L. Zheng, Y. Yang, Modality-invariant image-text embedding for image-sentence matching, ACM Transactions on Multimedia Computing, Communications, and Applications 15 (2019) 1–19.

Q. Zhang, Z. Lei, Z. Zhang, S. Z. Li, Context-aware attention network for image-text retrieval, in: CVPR, 2020, pp. 3536–3545.

Z. Liu, F. Chen, J. Xu, W. Pei, G. Lu, Image-text retrieval with cross-modal semantic importance consistency, IEEE Transactions on Circuits and Systems for Video Technology (2022).

L. Zhang, H. Wu, Q. Chen, Y. Deng, J. Siebert, Z. Li, Y. Han, D. Kong, Z. Cao, Vldeformer: Vision–language decomposed transformer for fast cross-modal retrieval, Knowledge-Based Systems 252 (2022) 109316.

Z. Li, C. Guo, X. Wang, H. Zhang, Y. Wang, Integrating listwise ranking into pairwise-based image-text retrieval, Knowledge-Based Systems 287 (2024) 111431.

C. Deng, X. Xu, H. Wang, M. Yang, D. Tao, Progressive cross-modal semantic network for zero-shot sketch-based image retrieval, IEEE Transactions on Image Processing 29 (2020) 8892–8902.

H. Wang, C. Deng, T. Liu, D. Tao, Transferable coupled network for zero-shot sketch-based image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1–1.

J. Li, Z. Ling, L. Niu, L. Zhang, Zero-shot sketch-based image retrieval with structure-aware asymmetric disentanglement, Computer Vision and Image Understanding 218 (2022) 103412.

S. Liang, W. Dai, Y. Cai, C. Xie, Sketch-based 3d shape retrieval via teacher–student learning, Computer Vision and Image Understanding 239 (2024) 103903.

Y. Yang, M. Wang, W. Zhou, H. Li, Cross-modal joint prediction and alignment for composed query image retrieval, in: ACM Multimedia, 2021, pp. 3303–3311.

G. Zhang, S. Wei, H. Pang, Y. Zhao, Heterogeneous feature fusion and cross-modal alignment for composed image retrieval, in: ACM Multimedia, 2021, pp. 5353–5362.

C. Gu, J. Bu, Z. Zhang, Z. Yu, D. Ma, W. Wang, Image search with text feedback by deep hierarchical attention mutual information maximization, in: ACM Multimedia, New York, NY, USA, 2021.

Y. Zhao, Y. Song, Q. Jin, Progressive learning for image retrieval with hybrid-modality queries, in: SIGIR, 2022, pp. 1012–1021.

X. Han, X. Zhu, L. Yu, L. Zhang, Y.-Z. Song, T. Xiang, Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks, in: CVPR, 2023, pp. 2669–2680.

M. Corbetta, G. L. Shulman, Control of goal-directed and stimulus-driven attention in the brain, Nature reviews neuroscience 3 (2002) 201–215.

A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the eleventh annual conference on Computational learning theory, 1998, pp. 92–100.

S. Qiao, W. Shen, Z. Zhang, B. Wang, A. Yuille, Deep co-training for semi-supervised image recognition, in: ECCV, 2018, pp. 135–152.

J. Peng, G. Estrada, M. Pedersoli, C. Desrosiers, Deep co-training for semi-supervised image segmentation, Pattern Recognition 107 (2020) 107269.

T. Hui, S. Liu, Z. Ding, S. Huang, G. Li, W. Wang, L. Liu, J. Han, Language-aware spatial-temporal collaboration for referring video segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: CVPR, 2018, pp. 3723–3732.

Z. Zheng, Y. Yang, Unsupervised scene adaptation with memory regularization in vivo, IJCAI (2019).

Y. Luo, L. Zheng, T. Guan, J. Yu, Y. Yang, Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation, in: CVPR, 2019, pp. 2507–2516.

T. L. Berg, A. C. Berg, J. Shih, Automatic attribute discovery and characterization from noisy web data, in: ECCV, 2010, pp. 663–676.

H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, R. Feris, Fashion iq: A new dataset towards retrieving images by natural language feedback, in: CVPR, 2021, pp. 11307–11317.

X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, L. S. Davis, Automatic spatially-aware fashion concept discovery, in: ICCV, 2017, pp. 1463–1471.

J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, B.-T. Zhang, Multimodal residual learning for visual qa, in: NeurIPS, volume 29, 2016.

E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, Film: Visual reasoning with a general conditioning layer, in: AAAI, volume 32, 2018.

G. Delmas, R. S. Rezende, G. Csurka, D. Larlus, Artemis: Attention-based retrieval with text-explicit matching and implicit similarity, in: ICLR, 2022.

Y. Chen, Z. Zheng, W. Ji, L. Qu, T.-S. Chua, Composed image retrieval with text feedback via multi-grained uncertainty regularization, arXiv:2211.07394 (2022).

A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv:1703.07737 (2017).

X. Wang, H. Zhang, W. Huang, M. R. Scott, Cross-batch memory for embedding learning, in: CVPR, 2020, pp. 6388–6397.

T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: ICML, 2020, pp. 1597–1607.

K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: CVPR, 2020, pp. 9729–9738.

S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

F. Zhang, M. Yan, J. Zhang, C. Xu, Comprehensive relationship reasoning for composed query based image retrieval, in: ACM Multimedia, 2022.

R. Girshick, Fast r-cnn, in: ICCV, 2015, pp. 1440–1448.

A. Graves, Long short-term memory, Supervised sequence labelling with recurrent neural networks (2012) 37–45.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv:1907.11692 (2019).

D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014).

J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, in: EMNLP-IJCNLP, 2019, pp. 6382–6388.