Comprehensive Linguistic-Visual Composition Network for Image Retrieval

Haokun Wen[†], Xuemeng Song[†*], Xin Yang[†], Yibing Zhan[§], Liqiang Nie[†*]

[†]Shandong University, Shandong, China, [§]JD Explore Academy, Beijing, China
{whenhaokun,sxmustc,joeyangbuer}@gmail.com,zhanyibing@jd.com,nieliqiang@gmail.com

ABSTRACT

Composing text and image for image retrieval (CTI-IR) is a new yet challenging task, for which the input query is not the conventional image or text but a composition, i.e., a reference image and its corresponding modification text. The key of CTI-IR lies in how to properly compose the multi-modal query to retrieve the target image. In a sense, pioneer studies mainly focus on composing the text with either the local visual descriptor or global feature of the reference image. However, they overlook the fact that the text modifications are indeed diverse, ranging from the concrete attribute changes, like "change it to long sleeves", to the abstract visual property adjustments, e.g., "change the style to professional". Thus, simply emphasizing the local or global feature of the reference image for the query composition is insufficient. In light of the above analysis, we propose a Comprehensive Linguistic-Visual Composition Network (CLVC-Net) for image retrieval. The core of CLVC-Net is that it designs two composition modules: fine-grained local-wise composition module and fine-grained global-wise composition module, targeting comprehensive multi-modal compositions. Additionally, a mutual enhancement module is designed to promote local-wise and global-wise composition processes by forcing them to share knowledge with each other. Extensive experiments conducted on three real-world datasets demonstrate the superiority of our CLVC-Net. We released the codes to benefit other researchers.

CCS CONCEPTS

• **Information systems** → **Image search**.

KEYWORDS

Linguistic-Visual Composition; Image Retrieval; Mutual Learning

ACM Reference Format:
Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, Liqiang Nie. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval.

*Xuemeng Song (sxmustc@gmail.com) and Liqiang Nie (nieliqiang@gmail.com) are corresponding authors.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462967

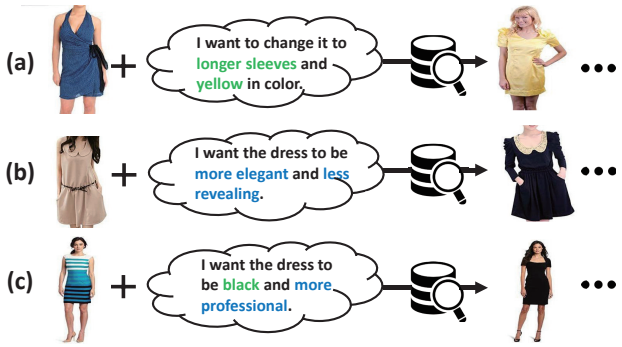

Figure 1: Three examples of composing text and image for image retrieval.

In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462967

1 INTRODUCTION

Image retrieval [9, 18, 19, 28, 29] refers to retrieving images that meet the user's search intent. Traditional image retrieval systems only allow users to use either the text or image query to express their search intent. However, in many cases, it is intractable for users to describe their search intent via a single textual query, meanwhile it is also difficult for users to find the ideal images to exactly convey their intent. Consequently, to allow users to flexibly express their search intent, composing text and image for image retrieval (CTI-IR) [37] is recently proposed and gaining increasing research attention.

As illustrated in Figure 1, the input of CTI-IR is a multi-modal query, i.e., a reference image plus a modification text. Notably, the text input is not the corresponding description of the given image but some modification intent on it. In light of this, the key to CTI-IR lies in how to properly compose the multi-modal query to retrieve the target image. According to the multi-modal composition manner, existing efforts can be broadly classified into two groups: *local-wise* [4, 16, 42] and *global-wise* [37] *composition methods*. The former focuses on composing the modification text with the local visual descriptors, e.g., feature maps, whereas the latter emphasizes the global feature of the reference image. Although existing researches have achieved promising results, they overlook the fact that the modifications are indeed diverse, spanning from concrete attribute changes to abstract visual property adjustments. We take the fashion-oriented image retrieval as an example, which is one of the most promising application scenarios for CTI-IR. Intuitively,

for the concrete attribute changes, like the modification need of *longer sleeves* in Figure 1(a), it is reasonable to give priority to the local representation of the reference image. In contrast, regarding abstract visual property modifications, like that in Figure 1(b), it seems that operating over the global representation of the reference image is more suitable. Meanwhile, Figure 1(c) shows the case where the text simultaneously contains both types of modifications. Overall, all the examples above suggest that simply utilizing the global or local representation of the reference image for query composition may lead to suboptimal performance. Motivated by this, we propose to incorporate both local-wise and global-wise compositions to better adapt to the diverse modification demands.

However, jointly modeling the local-wise and global-wise compositions for the task of CTI-IR is non-trivial due to the following two facts. 1) In most cases, there are only a few words or phrases that directly relate to the modification in the unstructured natural language text, like "*black*" and "*more professional*" in the case of Figure 1(c). Moreover, different words tend to refer to different regions of the reference image. Therefore, how to perform the fine-grained text-image composition from both local-wise and global-wise perspectives is a crucial challenge. And 2) although a given modification may be more suitable to be processed with either the local-wise or the global-wise composition, it in a sense inevitably involves compositions of both sides. For example, apart from the global abstract modification, "*less revealing*" also embraces some concrete local attribute changes, like extending the sleeve length and making the neckline higher. Moreover, since both composition manners correspond to the same target image, there should be certain latent consistency between the two composition ways. Consequently, how to seamlessly link both sides to take advantage of the underlying consistency between them forms a tough challenge.

To address the aforementioned challenges, we present a Comprehensive Linguistic-Visual Composition Network, dubbed as CLVC-Net, for image retrieval. As shown in Figure 2, CLVC-Net consists of four key modules: image/text encoding, fine-grained local-wise composition, fine-grained global-wise composition, and mutual enhancement. The first module works on extracting the intermediate representation of the image and text with two separate Convolution Neural Networks (CNNs) [24] and Long Short-Term Memory (LSTM) networks [15], respectively. The underlying philosophy of using two independent image/text encoders is to facilitate the following two split composition modules, where we argue that mingling these two types of compositions in one module may hurt the performance with an entangled optimization goal. The second and third modules of CLVC-Net devote to first capturing the fine-grained image-text alignments by corresponding attention modules, and then fulfilling the multi-modal compositions by respective affine transformations. Ultimately, the fourth module targets at distilling knowledge from one composition module to guide the other one in a mutual learning manner, where both the target ranking-level and the intermediate feature-level knowledge is extracted. Once CLVC-Net converges, the outputs of the two composition modules will be fused as the final query representation, which can be used for the target image retrieval.

Our main contributions can be summarized in three points:

- To the best of our knowledge, we are the first to unify the global-wise and local-wise compositions with mutual enhancement in the context of CTI-IR.
- We devise two affine transformation-based attentive composition modules, towards the fine-grained multi-modal compositions for both angles.
- Extensive experiments conducted on three real-world datasets validate the superiority of our model. As a byproduct, we released the codes to benefit other researchers[1].

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 details the proposed CLVC-Net. The experimental results and analyses are presented in Section 4, followed by the conclusion and future work in Section 5.

2 RELATED WORK

Our work is closely related to composing text and image for image retrieval (CTI-IR) and mutual learning.

2.1 CTI-IR

The early stage of CTI-IR [1, 41, 44] mainly focuses on the attribute manipulation of the reference image, where the modification text directly specifies the concrete attribute that needs to be manipulated. In particular, Zhao et al. [44] proposed a memory-augmented attribute manipulation network, where a memory block is introduced to store all the attribute template representations, and the corresponding attribute representation and the reference image representation will be fused to search the target image. Differently, Yang et al. [41] presented a generative attribute manipulation scheme for fashion retrieval, where a prototype image that meets the attribute manipulation requirements is synthesized by Generative Adversarial Networks to enhance the target item retrieval. Although previous studies have achieved remarkable success, they mainly restrict the user's modification to a set of pre-defined attributes, which limits their applications.

Towards this end, Vo et al. [37] proposed a new research field, composing natural language based text and image for image retrieval, which has drawn increasing research attention. Generally, according to the multi-modal fusion manner, existing efforts can be classified into two groups: local-wise and global-wise composition methods. The former [4, 16, 42] obtains the composed query representation by composing modification text representation with the local visual descriptor of the reference image. For example, to adaptively fulfil the multi-modal composition, VAL [4] fuses the text representation with the local feature maps of the reference image with an attention mechanism, and introduces the hierarchical matching regularization between the composed query representations and the target representations to enhance the retrieval performance. In contrast, the latter ones fulfil the modification over the global representation of the reference image. One typical example is TIRG [37], which fuses the global representation of the reference image and the text representation with a gated residual connection. Although these efforts have made prominent progress, facing the diverse modification needs that include both concrete attribute changes and abstract property adjustments, they are incompetent to achieve the optimal performance for CTI-IR.

[1] https://site2750.wixsite.com/clvcnet.

Notably, the recently proposed DCNet [22] incorporates both the local and global features of the reference image for composition. However, it simply cascades the global and local features to derive a more robust representation for the reference image. Beyond that, we design two split subnetworks to fulfil the fine-grained local-wise and global-wise compositions, respectively. Moreover, these two subnetworks are mutually enhanced by sharing knowledge to each other during the alternative optimization.

2.2 Mutual Learning

Knowledge distillation [14, 27, 35] is an effective and widely-used technique to transfer knowledge from a teacher network to a student network. This idea is first introduced by Hinton et al. [14] in the context of transferring knowledge from a large cumbersome model to a small one. Essentially, one key of knowledge distillation is the existence of the teacher network that possesses knowledge to guide the student network. However, in practice, there can be no explicit teacher but only students. Towards this, Zhang et al. [43] proposed the mutual learning, which aims to distill knowledge between students by pushing them to learn collaboratively and teach each other. Since then, mutual learning gets many researchers' attention. For example, Luo et al. [26] adopted mutual learning in person re-identification to boost model performance, where a set of student models are enforced to transfer knowledge to each other. In addition, Chan et al. [3] proposed the inconsistency loss for multi-task learning, which essentially shares the same loss function format with [43]. Inspired by these successful applications of mutual learning, in this work, we propose our mutual enhancement module to encourage the two composition modules in our model to learn collaboratively from both target ranking-level and feature-level and further boost the performance.

3 METHODOLOGY

In this section, we first formulate the problem and then detail the proposed CLVC-Net for image retrieval.

3.1 Problem Formulation

In this work, we aim to solve the CTI-IR problem, which can be formally defined as given a multi-modal query of a reference image and its modification text, we need to retrieve its corresponding target image from a set of gallery images. In light of this, it is essential to learn an accurate representation of the multi-modal query, i.e., a composed query, which can be used for the target image retrieval. Suppose that we have a set of triplets, denoted as $\mathcal{D} = \left\{(x_r, t_m, x_t)_i\right\}_{i=1}^{N}$, where x_r is the *reference image*, t_m is the *modification text*, x_t is the *target image*, and N is the total number of triplets. Based on $\mathcal{D}$, we aim to optimize a multi-modal composition scheme, which is able to learn the latent space where the representation of the multi-modal query (x_r, t_m) and that of the target image x_t should be close. Formally, we have,

$$\mathcal{H}(x_r, t_m) \rightarrow \mathcal{F}(x_t), \tag{1}$$

where $\mathcal{H}$ represents the transformation for mapping the multi-modal query to the latent space, while $\mathcal{F}$ denotes that for the target image.

3.2 CLVC-Net

As the major novelty, due to the concern that the text may refer to diverse modifications ranging from the concrete attribute changes to highly abstract visual property adjustments, we propose a Comprehensive Linguistic-Visual Composition Network (CLVC-Net) for image retrieval, as shown in Figure 2. It consists of four key modules: (a) image/text encoding, (b) fine-grained local-wise composition (FLC for short), (c) fine-grained global-wise composition (FGC for short), and (d) mutual enhancement. Firstly, the reference image and modification text are encoded by image encoder and text encoder, respectively (described in Section 3.2.1). Secondly, the intermediate representations are processed by the FLC and FGC modules to obtain the composed query representations, respectively (detailed in Sections 3.2.2 and 3.2.3). Last but not least, the FLC and FGC modules are mutually enhanced by sharing knowledge with each other (explained in Section 3.2.4). Once CLVC-Net gets converged, the outputs of the two composition modules will be fused as the final query representation for the target image retrieval. We now detail each module of our CLVC-Net.

3.2.1 Image/Text Encoding. As two modalities are involved, we first introduce the encoding of each modality.

Image Encoding. Regarding the image representation, we adopt the widely used CNNs, which have obtained remarkable success in many computer vision tasks [6, 8, 13, 17, 39, 40]. As to facilitate the final mutual enhancement between the FLC and FGC, we employ two separate CNNs as the image encoders for different composition perspectives, denoted as CNN^L and CNN^G, respectively. Specifically, the intermediate representations for the reference image and the target image can be given as follows,

$$\begin{cases} \mathbf{X}_r^L = \text{CNN}^L(x_r), \mathbf{X}_r^G = \text{CNN}^G(x_r), \\ \mathbf{X}_t^L = \text{CNN}^L(x_t), \mathbf{X}_t^G = \text{CNN}^G(x_t), \end{cases} \tag{2}$$

where $\mathbf{X}_r^L \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{X}_r^G \in \mathbb{R}^{C\times H\times W}$ refer to the reference image feature maps to be processed by the following FLC and FGC modules, respectively. While $\mathbf{X}_t^L \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{X}_t^G \in \mathbb{R}^{C\times H\times W}$ can be treated as the corresponding target ground truth representations. $C \times H \times W$ is the shape of feature maps.

Text Encoding. Similar to existing studies [4, 7, 37], to mine the sequential relationships among the words within the text, we use LSTMs as the text encoders[2]. Different from existing studies that mainly focus on the sentence-level text representation, we explore the word-level intermediate representations to facilitate the fine-grained image-text composition. Similar to the image encoding, we utilize two separate LSTMs, i.e., LSTM^L and LSTM^G, to get the text representations for FLC and FGC, respectively. Formally, the modification text representations can be obtained as follows,

$$\begin{cases} \mathbf{T}_m^L = [\mathbf{h}_1^L, \cdots, \mathbf{h}_U^L] = \text{LSTM}^L(t_m), \\ \mathbf{T}_m^G = [\mathbf{h}_1^G, \cdots, \mathbf{h}_U^G] = \text{LSTM}^G(t_m), \end{cases} \tag{3}$$

where $\mathbf{h}_i^L \in \mathbb{R}^D$ and $\mathbf{h}_i^G \in \mathbb{R}^D$ are the hidden vectors of the i-th word yielded by LSTM^L and LSTM^G, respectively. $\mathbf{T}_m^L \in \mathbb{R}^{D\times U}$

[2]Before feeding into the LSTM, the text is first tokenized into standard vocabularies.

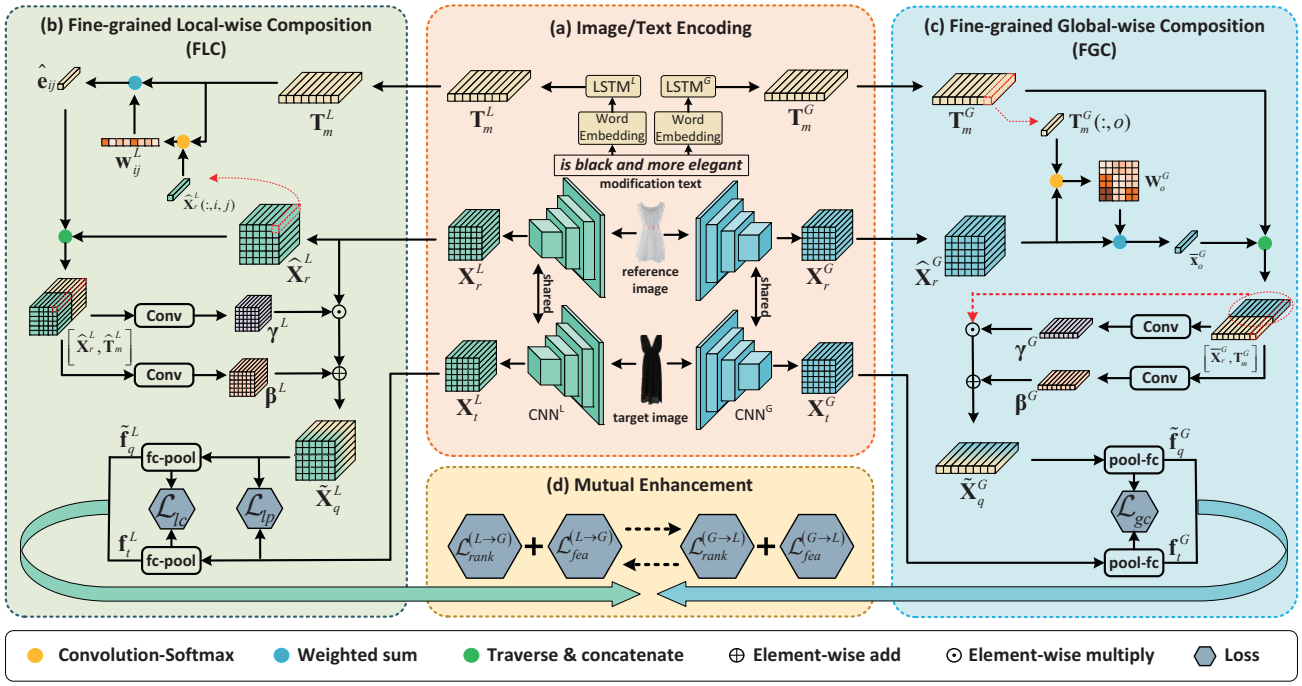

Figure 2: The proposed CLVC-Net consists of four key modules: (a) image/text encoding, (b) fine-grained local-wise composition (FLC), (c) fine-grained global-wise composition (FGC), and (d) mutual enhancement.

and $\mathbf{T}_m^G \in \mathbb{R}^{D\times U}$ are the text representations to be processed by the two composition modules. U is the number of words in the text.

3.2.2 Fine-grained Local-wise Composition (FLC). As aforementioned, the core task of CTI-IR can be formulated as learning the latent representation of the multi-modal query. In a sense, this goal can be achieved by transforming the reference image to the target image conditioned on the modification text. Therefore, to learn the latent representation of the multi-modal query, we resort to the affine transformation, which has been proven to be effective in many conditional image synthesis tasks [5, 25, 30]. Formally, the FLC can be formulated as follows,

$$\widetilde{\mathbf{X}}_q^L = \boldsymbol{\gamma}^L \odot \mathbf{X}_r^L + \boldsymbol{\beta}^L, \tag{4}$$

where $\widetilde{\mathbf{X}}_q^L \in \mathbb{R}^{C\times H\times W}$ denotes the local-wise composed query representation, $\boldsymbol{\gamma}^L \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{\beta}^L \in \mathbb{R}^{C\times H\times W}$ are the to-be-learned affine parameters, modulating the given reference image by scaling and shifting operations, respectively. $\odot$ denotes the Hadamard element-wise product.

Instead of directly extracting affine parameters from the textual representation like existing methods [25, 31], we argue that it is necessary to jointly consider the reference image and modification text. The underlying philosophy is rather straightforward as even for the same modification requirement, like changing the sleeve length, different reference images can involve different scaling or shifting operations. In other words, only by combining the text and image can we know where and how to perform the modification over the reference image. Moreover, unlike existing work [4] that concatenates the tiled representations of the modification text with the feature maps of the reference image to model the pair-wise image-text interaction, we introduce the attention mechanism to adaptively align the image and text representation. This is due to an intuitive fact that different spatial entries in the feature maps correspond to different textual semantics. Specifically, to facilitate the attention weight calculation, we first use the 1×1 convolution layer to transform each entry-wise feature map, i.e., $\mathbf{X}_r^L(:,i,j)$, into the same space of the modification text representation as follows,

$$\widehat{\mathbf{X}}_r^L = Conv1\left(\mathbf{X}_r^L\right), \tag{5}$$

where $\widehat{\mathbf{X}}_r^L \in \mathbb{R}^{D\times H\times W}$ denotes the transformed feature maps. Then, to derive the weight distribution over all words for each spatial entry, we perform the convolution operation over all word-level textual representations with each spatial-wise feature map, i.e., $\widehat{\mathbf{X}}_r^L(:,i,j)$, as follows,

$$\mathbf{w}_{ij}^L = \text{softmax}\left((\mathbf{T}_m^L \otimes \widehat{\mathbf{X}}_r^L(:,i,j))/\tau^L\right), \tag{6}$$

where $\otimes$ denotes the convolution operation. τ^L is the temperature factor, introduced to produce a softer weight distribution.

Based on the weight distribution vector $\mathbf{w}_{ij}^L \in \mathbb{R}^U$, we can get the weighted spatial-wise textual representation as follows,

$$\hat{\mathbf{e}}_{ij} = \sum_{\ell=1}^{U} \mathbf{w}_{ij}^L(\ell)\cdot \mathbf{T}_m^L(:,\ell), \tag{7}$$

where $\mathbf{w}_{ij}^L(\ell)$ is the ℓ-th element of $\mathbf{w}_{ij}^L$, indicating the importance of the ℓ-th word in the text towards the (i,j)-th feature map. For ease of illustration, we summarize all the weighted textual representations for all spatial entries with $\widehat{\mathbf{T}}_m^L \in \mathbb{R}^{D\times H\times W}$, where $\widehat{\mathbf{T}}_m^L(:,i,j) = \hat{\mathbf{e}}_{ij}$.

Ultimately, we concatenate the transformed image feature maps $\widehat{\mathbf{X}}_r^L$ and the weighted textual representations as $\widehat{\mathbf{T}}_m^L$ over the channel dimension. Thereafter, we feed the concatenated representations

to two convolutional networks to obtain the scaling and shifting parameters $\boldsymbol{\gamma}^L$ and $\boldsymbol{\beta}^L$ as follows,

$$\begin{cases} \boldsymbol{\gamma}^L = \mathcal{F}^L_{gamma}\left(\left[\widehat{\mathbf{X}}^L_r, \widehat{\mathbf{T}}^L_m\right]\right), \\ \boldsymbol{\beta}^L = \mathcal{F}^L_{beta}\left(\left[\widehat{\mathbf{X}}^L_r, \widehat{\mathbf{T}}^L_m\right]\right), \end{cases} \tag{8}$$

where convolutional networks $\mathcal{F}^L_{gamma}$ and $\mathcal{F}^L_{beta}$ are implemented with the inception convolution layers for their superior capability of representation learning [36].

Local-wise Metric Learning. To push the local-wise composed query one close to its target one in the latent space, we resort to the batch-based classification loss [37], which has shown compelling success in CTI-IR. Essentially, the batch-based classification loss encourages the ground truth target representation to be the closest one towards the composed query representation, while all the other candidates in the batch are treated as negative samples.

Specifically, we first transform the local-wise composed query representation and target representation of each triplet, i.e., $\widetilde{\mathbf{X}}^L_q$ and $\mathbf{X}^L_t$, into the vector forms, denoted as $\tilde{\mathbf{f}}^L_q$ and $\mathbf{f}^L_t$, with a pooling layer [32] followed by a fully-connected layer, respectively. Then according to the batch-based classification loss, we have,

$$\mathcal{L}_{lc} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{\cos\left(\tilde{\mathbf{f}}^L_{qi}, \mathbf{f}^L_{ti}\right)/\mu^L\right\}}{\sum_{j=1}^{B}\exp\left\{\cos\left(\tilde{\mathbf{f}}^L_{qi}, \mathbf{f}^L_{tj}\right)/\mu^L\right\}}\right\}, \tag{9}$$

where $\tilde{\mathbf{f}}^L_{qi}$ and $\mathbf{f}^L_{ti}$ stand for the local-wise composed query representation and the target representation of the i-th triplet sample in the batch, respectively. B is the batch size, and $\cos(\cdot,\cdot)$ denotes the cosine similarity function. μ^L is the temperature factor.

Due to the concern that the pooling operation may get some discriminative feature lost, we introduce an extra perceptual loss that is widely used in the neural style transfer [21] to further regulate the composed query feature maps to be close to its target. Specifically, the perceptual loss $\mathcal{L}_{lp}$ is defined as follows,

$$\mathcal{L}_{lp} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{C\times H\times W}\left\|\widetilde{\mathbf{X}}^L_{qi} - \mathbf{X}^L_{ti}\right\|_2^2, \tag{10}$$

where $\widetilde{\mathbf{X}}^L_{qi}$ and $\mathbf{X}^L_{ti}$ denote the intermediate feature maps of the composed query and that of its target image in the i-th triplet sample. $C\times H\times W$ is the feature map shape.

3.2.3 Fine-grained Global-wise Composition (FGC). FGC follows the similar paradigm with the FLC, except that, in this part we focus on the alignment between the modification text and the global representation of the reference image rather than the local feature maps. Intuitively, we can attach each word representation with the global vector of the reference image to explore the pair-wise text-image interaction and based on that to learn the scaling and shifting parameters for the FGC. However, this manner overlooks the fact that the sentence-based modification may involve multiple modification aspects, and each aspect can interact with different overviews of the reference image. As can be seen from Figure 1(b), the modification of *more elegant* may refer to a stylistic overview of the clothing, while *less revealing* may point to the overview pertaining to the garment length. Accordingly, beyond that, we propose to introduce an exclusive global feature vector for each word of the modification text, which is derived by adaptively summarizing the representations of all spatial entries.

Similar to the FLC, we first transform the feature maps $\mathbf{X}^G_r \in \mathbb{R}^{C\times H\times W}$ to $\widehat{\mathbf{X}}^G_r \in \mathbb{R}^{D\times H\times W}$ with a 1×1 convolution layer, to facilitate the weight map calculation. Then we convolve the image representation $\widehat{\mathbf{X}}^G_r$ with the o-th word representation at each spatial entry as follows,

$$\mathbf{W}^G_o = \text{softmax}\left((\widehat{\mathbf{X}}^G_r \otimes \mathbf{T}^G_m(:,o))/\tau^G\right), \tag{11}$$

where τ^G is the temperature factor for smoothing the weight distribution. Intuitively, each element of the weight map $\mathbf{W}^G_o$ refers to the importance of the corresponding spatial entry of the feature map towards the o-th modification word. Accordingly, we can derive the adaptive global representation $\overline{\mathbf{x}}^G_o$ for the o-th word as follows,

$$\overline{\mathbf{x}}^G_o = \sum_{i=1}^{W}\sum_{j=1}^{H}\mathbf{W}^G_o(i,j)\cdot\widehat{\mathbf{X}}^G_r(:,i,j). \tag{12}$$

Then we align each word representation with the corresponding global image representation, and derive the $\boldsymbol{\gamma}^G$ and $\boldsymbol{\beta}^G$ parameters for the global affine transformation as follows,

$$\begin{cases} \boldsymbol{\gamma}^G = \mathcal{F}^G_{gamma}\left(\left[\overline{\mathbf{X}}^G_r, \mathbf{T}^G_m\right]\right), \\ \boldsymbol{\beta}^G = \mathcal{F}^G_{beta}\left(\left[\overline{\mathbf{X}}^G_r, \mathbf{T}^G_m\right]\right), \end{cases} \tag{13}$$

where $\overline{\mathbf{X}}^G_r = [\overline{\mathbf{x}}^G_1, \overline{\mathbf{x}}^G_2, \cdots, \overline{\mathbf{x}}^G_U] \in \mathbb{R}^{D\times U}$. Similar to the FLC, both $\mathcal{F}^G_{gamma}$ and $\mathcal{F}^G_{beta}$ are implemented with the inception convolution layers. Ultimately, the FGC can be fulfilled by the following transformation,

$$\widetilde{\mathbf{X}}^G_q = \boldsymbol{\gamma}^G \odot \overline{\mathbf{X}}^G_r + \boldsymbol{\beta}^G, \tag{14}$$

where $\widetilde{\mathbf{X}}^G_q \in \mathbb{R}^{D\times U}$ stands for the global-wise composed query representation.

Global-wise Metric Learning. Similar to the local-wise metric learning, we also adopt the batch-based classification loss. Incorporating the sample index subscript, for clear illustration, the loss can be formulated as follows,

$$\mathcal{L}_{gc} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{\cos\left(\tilde{\mathbf{f}}^G_{qi}, \mathbf{f}^G_{ti}\right)/\mu^G\right\}}{\sum_{j=1}^{B}\exp\left\{\cos\left(\tilde{\mathbf{f}}^G_{qi}, \mathbf{f}^G_{tj}\right)/\mu^G\right\}}\right\}, \tag{15}$$

where $\tilde{\mathbf{f}}^G_{qi}$ and $\mathbf{f}^G_{ti}$ are the global-wise composed query vector and target vector for the i-th triplet sample, which can be derived from the corresponding global-wise composed query representation $\widetilde{\mathbf{X}}^G_{qi} \in \mathbb{R}^{D\times U}$ and target representation $\mathbf{X}^G_{ti} \in \mathbb{R}^{C\times H\times W}$ by the pooling layer [32] followed by a fully-connected layer, respectively.

3.2.4 Mutual Enhancement. As aforementioned in the Introduction, in a sense, although each modification may be more suitable to be handled by either the local-wise composition or the global-wise composition, it can still involve both the local modification and the global modification. As both composition manners refer to the same target image, there should be certain intrinsic consistency between the two composition modules. In light of this, we argue that if a given multi-modal query is easier to be processed by one composition manner, then the knowledge achieved by that

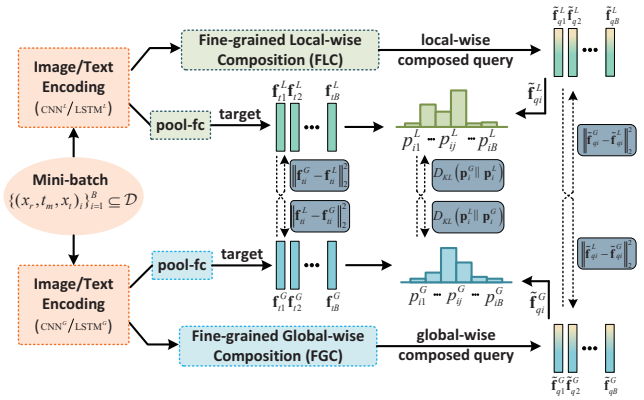

Figure 3: Mutual enhancement between two composition modules.

composition manner can be used for guiding the other one's learning. Propelled by this, instead of directly merging the losses of the FLC and FGC modules to optimize our CLVC-Net, we adopt the idea of mutual learning [43] and make the two modules alternatively share knowledge to each other. In particular, we enforce the two composition modules to mimic each other in terms of not only the final target ranking but also the intermediate features. Figure 3 illustrates the workflow of our mutual enhancement module.

On the one hand, we first take the by-product of the aforementioned batch-based classification loss, and get the normalized similarity between the i-th local-wise/global-wise composed query and j-th local-wise/global-wise target image as follows,

$$p_{ij}^{Z} = \frac{\exp\left\{\cos\left(\tilde{\mathbf{f}}_{qi}^{Z}, \mathbf{f}_{tj}^{Z}\right)/\nu^{Z}\right\}}{\sum_{j=1}^{B}\exp\left\{\cos\left(\tilde{\mathbf{f}}_{qi}^{Z}, \mathbf{f}_{tj}^{Z}\right)/\nu^{Z}\right\}}, Z \in \{L, G\}, \tag{16}$$

where ν^{Z} refers to the temperature factor. In this way, we can get the local-wise and global-wise batch-based target similarity distribution as $\mathbf{p}_i^L = [p_{i1}^L, p_{i2}^L, \cdots, p_{iB}^L]$ and $\mathbf{p}_i^G = [p_{i1}^G, p_{i2}^G, \cdots, p_{iB}^G]$, respectively. Then to encourage the ranking-level consistency between the two composition modules, we employ Kullback Leibler (KL) Divergence between $\mathbf{p}_i^L$ and $\mathbf{p}_i^G$. Specifically, due to the asymmetry of the KL divergence, we use $D_{KL}\left(\mathbf{p}_i^G \| \mathbf{p}_i^L\right)$ to optimize the local-wise network, while $D_{KL}\left(\mathbf{p}_i^L \| \mathbf{p}_i^G\right)$ is used to train the global-wise one. Taking the optimization of local-wise network as an example, we have,

$$\mathcal{L}_{rank}^{(G\to L)} = \frac{1}{B}\sum_{i=1}^{B} D_{KL}\left(\mathbf{p}_i^G \| \mathbf{p}_i^L\right) = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} p_{ij}^G \log\frac{p_{ij}^G}{p_{ij}^L}, \tag{17}$$

where $(G \to L)$ denotes the knowledge transferring from global-wise network to local-wise one. Notably, the optimization of global-wise network can be derived in a similar manner.

On the other hand, to promote knowledge transferring, we also introduce the feature-level consistency regularization, where we regulate the two composition modules to output consistent composed query representations and the target representations. Towards this end, $l2$ loss is utilized as follows,

$$\mathcal{L}_{fea}^{(G\to L)} = \mathcal{L}_{fea}^{(L\to G)} = \frac{1}{B}\sum_{i=1}^{B}\left(\left\|\tilde{\mathbf{f}}_{qi}^{G} - \tilde{\mathbf{f}}_{qi}^{L}\right\|_2^2 + \left\|\mathbf{f}_{ti}^{G} - \mathbf{f}_{ti}^{L}\right\|_2^2\right). \tag{18}$$

It is worth noting that although $\mathcal{L}_{fea}^{(G\to L)}$ and $\mathcal{L}_{fea}^{(L\to G)}$ share the same loss function, their optimization targets are different, where $\mathcal{L}_{fea}^{(G\to L)}$ and $\mathcal{L}_{fea}^{(L\to G)}$ aim to optimize the local-wise and global-wise networks, respectively.

Ultimately, we have the following objective function for optimizing the local-wise network,

$$\mathbf{\Theta}^{*} = \arg\min_{\mathbf{\Theta}}\left(\mathcal{L}_{lc} + \lambda\mathcal{L}_{lp} + \eta\left(\mathcal{L}_{rank}^{(G\to L)} + \mathcal{L}_{fea}^{(G\to L)}\right)\right), \tag{19}$$

where $\mathbf{\Theta}$ denotes the to-be-learned parameters in local-wise network, including CNN^L, LSTM^L, and FLC. λ and η are non-negative trade-off hyper-parameters. Similarly, the objective function for optimizing the global-wise network can be written as follows,

$$\mathbf{\Phi}^{*} = \arg\min_{\mathbf{\Phi}}\left(\mathcal{L}_{gc} + \eta\left(\mathcal{L}_{rank}^{(L\to G)} + \mathcal{L}_{fea}^{(L\to G)}\right)\right), \tag{20}$$

where $\mathbf{\Phi}$ denotes the to-be-learned parameters in global-wise network, including CNN^G, LSTM^G, and FGC. η is the trade-off hyper-parameter.

Notably, once our CLVC-Net is well-trained, we will rank the gallery images by jointly evaluating their cosine similarities to both local-wise and global-wise composed query representations, which are defined similarly as Eqns. (9) and (15), respectively.

4 EXPERIMENT

In this section, we first give the experimental settings and then detail the experiments conducted on three real-world datasets by answering the following research questions.

- **RQ1**: Does CLVC-Net surpass state-of-the-art methods?
- **RQ2**: How does each module affect CLVC-Net?
- **RQ3**: How is the quantitative performance of CLVC-Net?

4.1 Experimental Settings

4.1.1 Datasets. In the domain of CTI-IR, there have been several public datasets, including cube-oriented synthesized ones and fashion-oriented realistic ones. To evaluate the practical value of our model, we particular chose three real-world datasets: FashionIQ [11], Shoes [10], and Fashion200k [12].

FashionIQ [11] is a natural language-based interactive fashion retrieval dataset, crawled from *Amazon.com* and introduced for Fashion-IQ 2020 challenge[3]. It contains 77,684 fashion images covering three categories: Dresses, Tops&Tees, and Shirts. As a challenge dataset, only the training set of 18,000 triplets and the validation set of 6,016 triplets are available.

Shoes [2] is a dataset originally collected from *like.com* for the attribute discovery task, and developed by [10] with relative caption annotations for dialog-based interactive retrieval. The modification text in this dataset is also in the form of natural language. Following [4], we used 10k images for training and 4,658 images for evaluation.

Fashion200k [12] contains about 200K fashion images, each with an attribute-like text description. Following [4, 37], we only adopted the pair of images that has only one-word difference in their descriptions as the reference image and target image, while

[3]https://sites.google.com/view/cvcreative2020/fashion-iq.

Table 1: Performance comparison on FashionIQ and Shoes. † indicates the results are cited from [4], while ‡ denotes that from [22]. Best results are in boldface, while second best results are underlined.

Method	FashionIQ								Shoes		
	Dress		Shirt		Tops&Tees		Avg				
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R@1	R@10	R@50
Image_Only	4.96	13.34	5.40	14.13	5.10	13.26	5.15	13.58	7.38	33.23	58.73
Text_Only	6.89	21.67	8.54	26.25	8.82	27.64	8.08	25.19	0.57	5.96	18.63
Concatenation	9.97	26.77	9.37	28.07	7.75	25.09	9.03	26.64	6.31	31.92	58.62
Relationship (Santoro et al. 2017)†	15.44	38.08	18.33	38.63	21.10	44.77	18.29	40.49	12.31	45.10	71.45
Film (Perez et al. 2018)†	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04	10.19	38.89	68.30
TIRG (Vo et al. 2019)†	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	12.60	45.45	69.39
VAL (Chen et al. 2020)†	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04	16.49	49.12	73.53
DCNet (Kim et al. 2021)‡	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89	-	53.82	79.33
CLVC-Net	**29.85**	**56.47**	**28.75**	**54.76**	**33.50**	**64.00**	**30.70**	**58.41**	**17.64**	**54.39**	**79.47**

the modification text is synthesized by templates, such as "replace *red* with *green*". Same as [4, 37], we obtained around 172k triplets for training and 33k triplets for evaluation.

4.1.2 Implementation Details. For the image encoder, we selected ResNet50 [13] as the backbone, and discarded the down-sampling between the Stage 3 and Stage 4 [38] to preserve more detailed information in the feature map. Accordingly, the intermediate representations of reference/target images have the shape of $2048 \times 14 \times 14$. Pertaining to the text encoder, we set the dimension of the hidden layer in LSTM to 1024. In addition, we set the dimensions of the composed query representations and target representations of both composition modules as 1024. Regarding the local-wise affine transformation parameters learning, we implemented the inception convolution layers in Eqn.(8) with a convolution layer, which is split into four branches with different receptive fields as in [36], and followed by a batch normalization layer [20] and Relu activation function. Those in the FGC follow the same structure, except that they are constructed by 1D convolution manner.

We alternatively trained the two compositions by Adam optimizer [23] with an initial learning rate of 0.0001, which multiplies 0.1 at the 10-th epoch. We empirically set the batch size as 32, trade-off hyper-parameters in Eqn.(19) and Eqn.(20) as $\lambda = \eta = 1$. Temperature factors τ^L and τ^G in Eqn.(5) and Eqn.(11) are set to 7.0 and 4.0, respectively, while others, i.e., μ^L, μ^G, ν^L, and ν^G, are set to 10.0. For a fair comparison, dataset settings and evaluation metrics are kept the same as previous efforts [4, 22]. We utilized recall at rank k ($R@k$) that measures the fraction of queries for which the ground truth target is retrieved among the top k results. All the experiments are implemented by PyTorch, and we fixed the random seeds to ensure the reproducibility.

4.2 On Model Comparison (RQ1)

To validate the effectiveness of our method in the context of CTI-IR, we chose the following baselines, including both naive and state-of-the-art methods, for comparison.

- **Image_Only** simply takes the reference image representation as the composed query representation.
- **Text_Only** simply takes the modification text representation as the composed query representation.
- **Concatenation** feeds the concatenation of the reference image and modification text representations to a two-layer MLP to obtain the composed query representation.
- **Relationship** [34] summarizes the composed query representation based on a set of relational features, which are obtained by concatenating the textual features and a pair of local-wise visual features.
- **Film** [31] uses affine transformation to inject the text information into the feature maps of the reference image, and based on which to retrieve the target image.
- **TIRG** [37] resorts to gating mechanism to adaptively preserve and transform the reference image to get the composed query representation.
- **VAL** [4] utilizes the attention mechanism to achieve the preservation and transformation of the reference image. For fair comparison, we referred its performance when no extra textual description of the reference image is used.
- **DCNet** [22] wins the first place in Fashion-IQ 2020 challenge by jointly modeling the multi-granularity features of the reference image and modification text with a refined version of TIRG [37], and introduces a correction network on the difference between the reference image and target image.
- **JGAN** [42] utilizes a graph attention network to adaptively compose the modification text and reference image, where the local features of the reference image are extracted by Faster R-CNN model [33].
- **LBF** [16] fulfills the task with a cross-modal attention module, which is able to compose the local visual features of the reference image and the word-level representations of the modification text.

Tables 1 and 2 show the performance comparison among different methods on the three datasets, while for JGAN [42] and LBF [16], we only reported their performance on Fashion200k, since they did not publish their results on the other two datasets. From these tables, we obtained the following observations. 1) CLVC-Net consistently outperforms all baseline methods over all datasets. This confirms the advantage of our model that incorporates both the local-wise and global-wise composition in the context of CTI-IR. 2) The naive methods, i.e., Image_Only, Text_Only, and Concatenation, perform worse than the other methods, demonstrating the necessity of a proper query composition. And 3) DCNet achieves the second

Table 2: Performance comparison on Fashion200k. † and ‡ denote the results are cited from [4] and their own papers, respectively. Best results are in boldface, while second best are underlined.

Method	R@1	R@10	R@ 50
Image_Only	3.7	18.9	37.3
Text_Only	1.8	12.5	22.5
Concatenation	9.8	33.0	52.7
Relationship (Santoro et al. 2017)†	13.0	40.5	62.4
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)†	<u>21.2</u>	<u>49.0</u>	<u>68.8</u>
DCNet (Kim et al. 2021)‡	-	46.9	67.6
JGAN (Zhang et al. 2020)‡	17.3	45.3	65.7
LBF (Hosseinzadeh et al. 2020)‡	17.8	48.4	68.5
CLVC-Net	**22.6**	**53.0**	**72.2**

best results on FashionIQ and Shoes, while VAL ranks the second on Fashion200k, which to some extent implies that their methods cannot meet the various modification demands well across different datasets and suffer from the limited generalization ability. This further reflects the superiority of our CLVC-Net, which is able to cater for various modification demands.

4.3 On Ablation Study (RQ2)

To verify the importance of each module in our model, we also compared CLVC-Net with its following derivatives.

- **w/ Avgpool**: To study the effect of the fine-grained multi-modal composition strategy, we replaced the attention mechanism with the simple average pooling to get $\widehat{\mathbf{T}}_m^L$ and $\overline{\mathbf{X}}_r^G$, leaving all words or spatial entries being concatenated with the same visual/textual representation.
- **w/o Mutual**: To explore the effect of the mutual enhancement, we removed the knowledge distillation between two composition modules by setting η=0.
- **w/o Mutual-ranking**: To investigate the effect of the ranking-level knowledge in mutual enhancement, we removed the KL Divergence losses $\mathcal{L}_{rank}^{(G\to L)}$ and $\mathcal{L}_{rank}^{(L\to G)}$.
- **w/o Mutual-feature**: To get more insight into the feature-level knowledge in mutual enhancement, we removed the $l2$ losses $\mathcal{L}_{fea}^{(G\to L)}$ and $\mathcal{L}_{fea}^{(L\to G)}$.
- **Local-wise_Only** and **Global-wise_Only**: To check the importance of both composition modules, we removed each module from the training framework, respectively.
- **Local-wise_Only+** and **Global-wise_Only+**: To verify whether the two composition modules are able to absorb knowledge from each other, we trained the network with mutual enhancement, but only used the local-wise/global-wise composed query representation for retrieval.

Table 3 shows the ablation results of our CLVC-Net. From this table, we gained the following observations. 1) w/ Avgpool performs worse than our CLVC-Net, which proves the necessity of incorporating the fine-grained multi-modal composition. 2) CLVC-Net surpasses w/o Mutual-ranking, indicating that mutual enhancement is indeed helpful for integrating the two composition modules. 3) Both w/o Mutual-ranking and w/o Mutual-feature are

Table 3: Ablation study on FashionIQ and Shoes. The results on FashionIQ are the average result of three categories.

Method	FashionIQ (Avg)		Shoes	
	R@10	R@50	R@10	R@50
w/ Avgpool	29.10	57.06	53.05	78.02
w/o Mutual	28.81	56.25	51.86	77.11
w/o Mutual-kl	29.06	57.07	53.42	78.56
w/o Mutual-feature	28.94	56.34	53.05	78.05
Local-wise_Only	24.67	52.17	47.49	75.01
Global-wise_Only	25.20	50.91	47.80	73.10
Local-wise_Only+	28.19	56.06	52.26	78.30
Global-wise_Only+	29.51	57.31	53.34	78.33
CLVC-Net	**30.70**	**58.41**	**54.39**	**79.47**

superior to w/o Mutual, which suggests that it is essential to consider both ranking-level and feature-level knowledge in the mutual enhancement to gain the better knowledge transferring. 4) Compared to other derivatives, Local-wise_Only and Global-wise_Only present the worst performance, demonstrating that using only local-wise or global-wise composition module is suboptimal for CTI-IR. And 5) Local-wise_Only+ and Global-wise_Only+ significantly outperform Local-wise_Only and Global-wise_Only. This reconfirms the benefit of making the two composition modules share knowledge with each other.

To gain a deeper understanding of the superiority of our CLVC-Net over Local-wise_Only and Global-wise_Only, we visualized the word cloud based on the modification text of testing samples that are correctly retrieved at the top one place by different methods. Notably, to avoid the frequency bias, we normalized the frequency of each modification word in the correctly retrieved samples by its overall frequency in the test set. From Figure 4, we observed that as compared to abstract visual property changes, Local-wise_Only performs better at tackling the concrete attributes modifications, such as "bule" and "sleeves" on FashionIQ, and "closures" on Shoes. Conversely, Global-wise_Only shows superiority in processing the abstract visual property changes, like the style word "American" on FashionIQ and "style" on Shoes, rather than concrete attributes modifications. Meanwhile, we noticed our CLVC-Net performs well in both concrete modifications, e.g., "strips" on FashionIQ and "closures" on Shoes, and abstract modifications, e.g., "deeper" on FashionIQ and "style" on Shoes. This indicates that beyond Local-wise_Only and Global-wise_Only, our CLVC-Net is capable of handling diverse modifications, which proves the necessity of incorporating both local-wise and global-wise compositions in solving the CTI-IR task again.

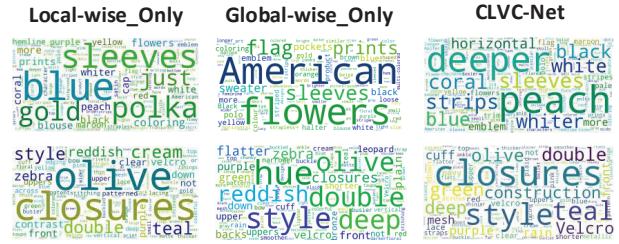

Figure 4: Comparison of word cloud on (a) FashionIQ and (b) Shoes.

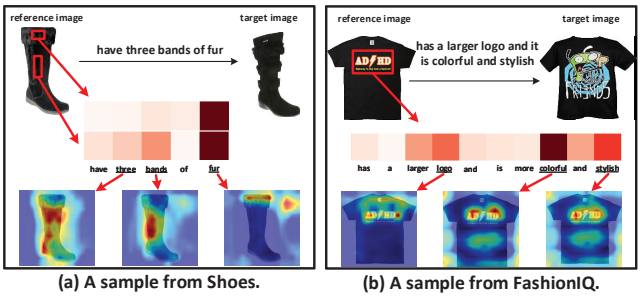

Figure 5: Attention visualization.

4.4 On Case Study (RQ3)

4.4.1 Attention Visualization. We visualized the attention mechanism in the FLC and FGC with two testing samples in Figure 5, where both the word weight distribution in FLC and spatial weight map in FGC are provided. Notably, to provide the more meaningful word weight calculation, we visualized the average word weight distribution of the corresponding modification region rather than a single spatial entry point. Meanwhile, for ease of illustration, we presented the spatial weight maps for the most informative modification words. As can be seen from Figure 5(a), the top part of the boot pays more attention to the word "fur", while the middle part of the boot further emphasized words "three" and "bands". Checking the reference and target images, we found the word weight distributions are reasonable. Meanwhile, we observed that the word "three" and "bands" are most related to the middle region of the boot, while "fur" concerns the top of the boot most, which part is does made of fur. Regarding that in Figure 5(b), the "larger logo", "colorful", and "stylish" are identified as the most informative words corresponding to the middle top part of the T-shirt. Furthermore, the word "logo" focuses on the pattern part of the T-shirt, while the words "colorful" and "stylish" emphasize nearly the whole T-shirt. Combining with the reference and target images, these observations are also meaningful. Overall, we can confirm the effectiveness of CLVC-Net in capturing the fine-grained text-image alignment.

4.4.2 Image Retrieval Visualization. Figure 6 illustrates several CTI-IR results obtained by our CLVC-Net on three datasets. Due to the limited space, we reported the top 5 retrieved images, and used blue and green boxes to denote the reference and target images, respectively. As shown in Figure 6(a), the target images are ranked at the first places, verifying that CLVC-Net could handle the single attribute modification well. Regarding the cases of FashionIQ and Shoes, where the modification text is human-written, and involves multiple modification aspects including both concrete attributes and abstract visual properties, our model still work well for some cases, as shown in the first row of Figure 6(b) and Figure 6(c). Meanwhile, we also noticed some failing examples, where the target images are not ranked at the top places. For example, in the second row in Figure 6(b), CLVC-Net misses the target image in the top 5 retrieved images. Nevertheless, checking the ranking list, we noticed that all the top 5 retrieved items meet the requirement of the modification text over the reference image. As for the second example in Figure 6(c), the target image with the "tiger print" is ranked fourth behind those with "snake print" and "leopard print". This may be due to the fact that these features are so rare in the training set that the model fails to fully distinguish them. Overall, these observations verify the practical value of CLVC-Net.

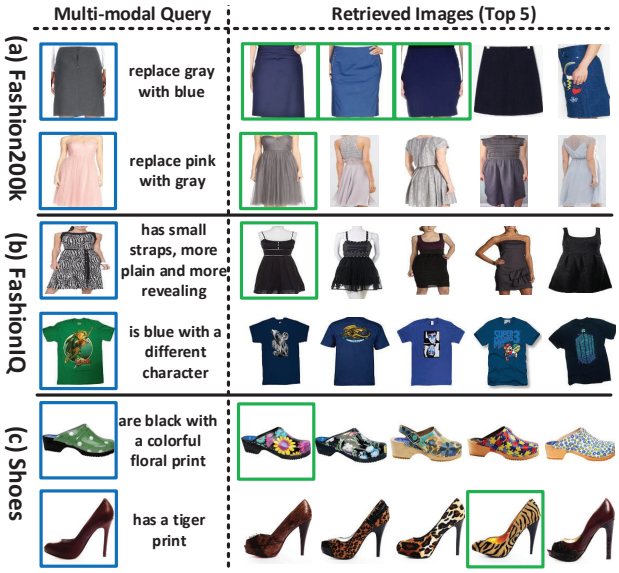

Figure 6: Illustration of several CTI-IR results obtained by our CLVC-Net on three datasets.

5 CONCLUSION AND FUTURE WORK

In this work, we present a novel comprehensive linguistic-visual composition network to tackle the challenging CTI-IR task, which seamlessly unifies the fine-grained local-wise composition and fine-grained global-wise composition with mutual enhancement. In particular, we propose two affine transformation-based attentive composition modules, corresponding to the two compositions, respectively. Moreover, to capture the underlying consistency between the two compositions, we introduce the mutual enhancement strategy to make the two compositions share knowledge with each other. Extensive experiments have been conducted on three public datasets, and the results demonstrate the effectiveness of our method. Furthermore, to gain a deep insight into our approach, we performed sufficient ablation studies, and visualized the case studies. As expected, we found that the local-wise composition manner does well in concrete attribute changes, while the global-wise one is adept at abstract visual property adjustments. Nevertheless, both of them cannot meet the diverse modification demands well, comfirming the necessity of simultaneously incorporating the two compositions. Additionally, we noticed that using mutual enhancement can significantly boost each composition module's performance, which confirms that the knowledge mutually transferred is helpful to CTI-IR. To take it further, we will extend our method to solve the multi-turn interactive image retrieval task, which is an essential problem in the multimodal dialogue systems.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.:U1936203; the Key R&D Program of Shandong (Major scientific and technological innovation projects), No.:2020CXGC010111; new AI project towards the integration of education and industry in QLUT.

REFERENCES

[1] Kenan E. Ak, Ashraf A. Kassim, Joo-Hwee Lim, and Jo Yew Tham. 2018. Learning Attribute Representations with Localization for Flexible Fashion Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7708–7717.

[2] Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. 2010. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *Proceedings of the European Conference on Computer Vision*. Springer, 663–676.

[3] Hou Pong Chan, Wang Chen, and Irwin King. 2020. A Unified Dual-view Model for Review Summarization and Sentiment Classification with Inconsistency Loss. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1191–1200.

[4] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image Search with Text Feedback by Visiolinguistic Attention Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2998–3008.

[5] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. 2017. Semantic Image Synthesis via Adversarial Learning. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5707–5715.

[6] Fuli Feng, Xiangnan He, Jie Tang, and Tat-Seng Chua. 2019. Graph Adversarial Training: Dynamically Regularizing Based on Graph Structure. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[7] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. Temporal Relational Ranking for Stock Prediction. *ACM Transactions on Information Systems* 37, 2 (2019), 1–30.

[8] Fuli Feng, Weiran Huang, Xiangnan He, Xin Xin, Qifan Wang, and Tat-Seng Chua. 2021. Should Graph Convolution Trust Neighbors? A Simple Causal Inference Method. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1–11.

[9] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2251–2260.

[10] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogério Schmidt Feris. 2018. Dialog-based Interactive Image Retrieval. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 676–686.

[11] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogério Schmidt Feris. 2019. The Fashion IQ Dataset: Retrieving Images by Combining Side Information and Relative Natural Language Feedback. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1–14.

[12] 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*. IEEE, 1472–1480.

[13] 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*. IEEE, 770–778.

[14] Geoffrey Hinton, Jeff Dean, and Oriol Vinyals. 2014. Distilling the Knowledge in a Neural Network. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 1–9.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[16] Mehrdad Hosseinzadeh and Yang Wang. 2020. Composed Query Image Retrieval Using Locally Bounded Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3593–3602.

[17] Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. 2021. Video Moment Localization via Deep Cross-modal Hashing. *IEEE Transactions on Image Processing* 30 (2021), 4667–4677.

[18] Yupeng Hu, Peng Zhan, Yang Xu, Jia Zhao, Yujun Li, and Xueqing Li. 2021. Temporal Representation Learning for Time Series Classification. *Neural Computing and Applications* 33, 8 (2021), 3169–3182.

[19] Fei Huang, Yong Cheng, Cheng Jin, Yuejie Zhang, and Tao Zhang. 2017. Deep Multimodal Embedding Model for Fine-grained Sketch-based Image Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 929–932.

[20] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the International Conference on Machine Learning*. ACM, 448–456.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proceedings of the European Conference on Computer Vision*. Springer, 694–711.

[22] 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*. AAAI, 1–9.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 1–15.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 1106–1114.

[25] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. 2020. ManiGAN: Text-Guided Image Manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7877–7886.

[26] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. 2019. AlignedReID++: Dynamically Matching Local Information for Person Re-identification. *Pattern Recognition* 94 (2019), 53–61.

[27] Lingjuan Lyu and Chi-Hua Chen. 2020. Differentially Private Knowledge Distillation for Mobile Analytics. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1809–1812.

[28] Changyi Ma, Chonglin Gu, Wenye Li, and Shuguang Cui. 2020. Large-scale Image Retrieval with Sparse Binary Projections. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1817–1820.

[29] David Novak, Michal Batko, and Pavel Zezula. 2015. Large-scale Image Retrieval using Neural Net Descriptors. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1039–1040.

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2337–2346.

[31] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 3942–3951.

[32] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 7 (2019), 1655–1668.

[33] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 91–99.

[34] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A Simple Neural Network Module for Relational Reasoning. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 4967–4976.

[35] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. 2018. Neural Compatibility Modeling with Attentive Knowledge Distillation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 5–14.

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–9.

[37] 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 Conference on Computer Vision and Pattern Recognition*. IEEE, 6439–6448.

[38] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 274–282.

[39] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1437–1445.

[40] Xin Yang, Xuemeng Song, Fuli Feng, Haokun Wen, Ling-Yu Duan, and Liqiang Nie. 2021. Attribute-wise Explainable Fashion Compatibility Modeling. *ACM Transactions on Multimedia Computing, Communications and Application* 17, 1 (2021), 36:1–36:21.

[41] Xin Yang, Xuemeng Song, Xianjing Han, Haokun Wen, Jie Nie, and Liqiang Nie. 2020. Generative Attribute Manipulation Scheme for Flexible Fashion Search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 941–950.

[42] Feifei Zhang, Mingliang Xu, Qirong Mao, and Changsheng Xu. 2020. Joint Attribute Manipulation and Modality Alignment Learning for Composing Text and Image to Image Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 3367–3376.

[43] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep Mutual Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4320–4328.

[44] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6156–6164.