CLIP-Based Composed Image Retrieval with Comprehensive Fusion and Data Augmentation

Haoqiang Lin[1], Haokun Wen[2], Xiaolin Chen[1], and Xuemeng Song[1](✉)

[1] Shandong University, Shandong, China
zichaohq@gmail.com, cxlicd@gmail.com, sxmustc@gmail.com
[2] Harbin Institute of Technology (Shenzhen), Guangdong, China
whenhaokun@gmail.com

Abstract. Composed image retrieval (CIR) is a challenging task where the input query consists of a reference image and its corresponding modification text. Recent methodologies harness the prowess of visual-language pre-training models, *i.e.*, CLIP, yielding commendable performance in CIR. Despite their promise, several shortcomings linger. First, a salient domain discrepancy between the CLIP's pre-training data and the CIR's training data leads to suboptimal feature representation. Second, the existing multimodal fusion mechanisms solely rely on weighted summing and feature concatenation, neglecting the intricate higher-order interactions inherent in the multimodal query. This oversight poses challenges in modeling complex modification intents. Additionally, the paucity of data impedes model generalization. To address these issues, we propose a CLIP-based composed image retrieval model with comprehensive fusion and data augmentation (CLIP-CD), consisting of two training stages. In the first stage, we fine-tune both the image and text encoders of CLIP to alleviate the aforementioned domain discrepancy. In the second stage, we propose a comprehensive multimodal fusion module that enables the model to discern complex modification intentions. Furthermore, we propose a similarity-based data augmentation method for CIR, ameliorating data scarcity and enhancing the model's generalization ability. Experimental results on the Fashion-IQ dataset demonstrate the effectiveness of our method.

Keywords: Image retrieval · Vision-Language pre-training model · Multimodal fusion

1 Introduction

Image retrieval [7] stands as a cornerstone within the computer vision field, playing pivotal roles in diverse domains ranging from face recognition [19] to fashion retrieval [27]. Traditional image retrieval has predominantly centered on single-modal queries, including text-based image retrieval [22] and content-based image retrieval [17]. However, in many cases, expressing precise search intent via a single-modal query often poses formidable challenges for users.

T. Liu et al. (Eds.): AI 2023, LNAI 14471, pp. 190–202, 2024.
https://doi.org/10.1007/978-981-99-8388-9_16

To address the limitations of conventional image retrieval, composed image retrieval (CIR) [21] has been proposed and gained increasing research attention. In this task, the input is a multimodal query, *i.e.*, a reference image plus a modification text. The reference image reflects the user's overarching retrieval demands, while the modification text delineates the user's specific unsatisfactory features in the reference image and his/her desired modifications. This multimodal query enables users to express their retrieval intents more flexibly and accurately.

Recent efforts have been increasingly devoted toward CIR. Predominantly, these works apply conventional frameworks like CNNs [12] and LSTM [8] to cultivate representations for the multimodal query. Yet, with the burgeoning prowess of vision-language pre-training models in feature extraction, the innovative method Clip4Cir [2] has integrated CLIP [4] and achieved impressive results. Nevertheless, there are still some limitations that need to be addressed. 1) Clip4Cir overlooks the substantial domain discrepancy between the CLIP's pre-training image data and the CIR's image data, which results in suboptimal feature extraction. 2) Clip4Cir employs a simple combiner network reliant on mere weighted summing and feature concatenation. It neglects the higher-order interaction between the multimodal query and potentially fails to model complex modification intents. And 3) the laborious data annotation limits the scale of most CIR's datasets. Like other existing works, Clip4Cir overlooks this issue, resulting in insufficient model generalization.

To address the above limitations, we present a CLIP-based composed image retrieval model with comprehensive fusion and data augmentation (CLIP-CD), as illustrated in Fig. 1, which comprises two stages. In the first stage, we fine-tune the CLIP's image and text encoders to alleviate the domain discrepancy problem. In the second stage, we design a multimodal fusion module, which incorporates weighted summing, feature concatenation, and bilinear pooling [20], to enhance the model's multimodal fusion capabilities. Moreover, we propose a similarity-based data augmentation method to expand the dataset size and enhance the model's generalization capabilities. Extensive experiments on the Fashion-IQ dataset corroborate the superiority of our method.

Our main contributions can be summarized as follows:

- We present a novel CLIP-based method for CIR. Our approach not only incorporates a fine-tuning strategy specifically designed to address the domain discrepancy problem but also integrates a comprehensive multimodal fusion module to enhance the effectiveness of multimodal fusion.
- To the best of our knowledge, we are the first to introduce a similarity-based data augmentation mechanism in CIR, which alleviates the insufficient training data problem.
- Extensive experiments conducted on the real-world Fashion-IQ dataset validate the superiority of our model.

2 Related Work

Our work is closely related to composed image retrieval (CIR) and vision-language pre-training (VLP).

2.1 Composed Image Retrieval

Recently, there have been numerous works aiming to solve this problem. For example, Vo *et al.* [21] employed gate mechanisms coupled with residual modules to fuse the multimodal query features. Later, Lee *et al.* [13] handled changes in both content and style conveyed by modification text through the designed content modulator and the style modulator. Meanwhile, Wen *et al.* [23] harnessed the mutual learning strategy to unify both local-wise multimodal fusion and global-wise multimodal fusion. Baldrati *et al.* [2] pioneered the integration of CLIP into this task and achieved remarkable performance. However, they overlooked the domain discrepancy between the CLIP's pre-training image data and the CIR's image data. The combiner network employed in their approach also exhibits limitations in effectively modeling complex modification intents within multimodal queries. Furthermore, like other works, they fail to address the issue of the limited scale of most CIR's datasets, resulting in constrained model generalization. In light of this, we fine-tuned both the image and text encoders to alleviate the domain discrepancy, and also introduced a comprehensive multimodal fusion module to effectively capture complex modification intents. Besides, we proposed a similarity-based data augmentation method to expand the dataset size and enhance the model's generalization capabilities.

2.2 Vision-Language Pre-training

Vision-language pre-training models leverage vast data for pre-training and generalize well on numerous downstream tasks through fine-tuning. Examples of vision-language pre-training models include ViLBERT [16], Oscar [14], and CLIP. Among them, CLIP stands out due to its contrastive learning based on 400 million image-text pairs. It can handle both text and visual inputs and model the relationship between them. This capability has led to advancements in multimodal areas like fine-grained classification [4], zero-shot retrieval [5], and visual commonsense reasoning [22]. Drawing from these insights, we proposed an effective fine-tuning strategy to bridge the gap between the CLIP's pre-training data and the CIR's data.

3 Methodology

In this section, we first formulate the problem, and then detail the proposed CLIP-CD.

3.1 Problem Formulation

In this work, we aim to tackle the CIR task, which can be formally defined as that given a multimodal query comprising a reference image and its modification text, the goal is to retrieve the optimal target image from a set of gallery images. Suppose we have a set of triples denoted as $\mathcal{D} = \{(I_r, T_m, I_t)_i\}_{i=1}^{N}$, where I_r is the reference image, T_m is the modification text, I_t signifies the target image, and N

is the total number of triplets. Based on $\mathcal{D}$, our goal is to train a model that can effectively fuse the multimodal query (I_r, T_m) to be close to the representation of the target image I_t. This can be formalized as follows,

$$f(I_r, T_m) \rightarrow h(I_t), \tag{1}$$

where $f(\cdot)$ represents the multimodal fusion function mapping the multimodal query to the latent space, while $h(\cdot)$ denotes the feature embedding function for the target image.

3.2 CLIP-CD

As illustrated in Fig. 1, we propose a CLIP-based composed image retrieval model with comprehensive fusion and data augmentation, which consists of two training stages. In the first stage (encoder fine-tuning stage), we simultaneously fine-tune the text and image encoders of CLIP, which helps alleviate the problem of domain discrepancy. In the second stage (multimodal fusion stage), we freeze the parameters of the CLIP's encoder fine-tuned in the first stage and focus on learning a multimodal fusion module. Additionally, to further expand the dataset size and improve the model's generalization, we generate pseudo triplets by replacing the reference/target image with another similar one.

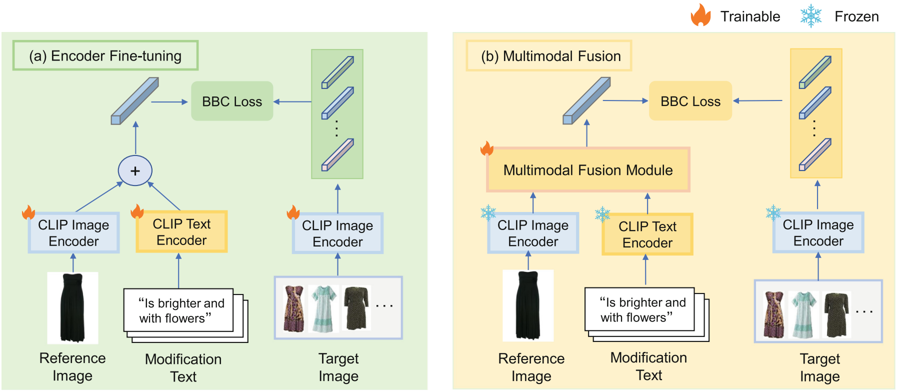

Fig. 1. The overall architecture of the proposed framework consists of two training stages: (a) Encoder Fine-tuning and (b) Multimodal Fusion. The parameters of the image encoder are shared by the reference image and the target image.

3.2.1 Encoder Fine-Tuning

The first step is to address the domain discrepancy between the data of the CLIP's pre-training and that of the CIR's training.

Regarding text data, CLIP is trained on descriptive texts directly related to images, such as "A photo of a dog". However, CIR's modification texts highlight differences between the reference and target images, such as "has more colors and is purple". In addition, as for image data, CLIP's pre-training data contain images from diverse domains in the open domain, like objects, landscapes, and humans. Conversely, CIR's training image data are usually domain-specific, such as fashion-related items. Hence, there are significant domain discrepancies between these two tasks in both image and text training data.

Based on the above analysis, in this stage, we focus on fine-tuning both the image and text encoders of CLIP to address the domain discrepancy problem. Figure 1(a) illustrates the overview of the first stage. Specifically, we first utilize CLIP to extract image and text features from the training triplet (I_r, T_m, I_t), which can be formulated as follows,

$$\begin{cases} \mathbf{x_r} = \mathrm{IE}\left(I_r\right), \\ \mathbf{t_m} = \mathrm{TE}\left(T_m\right), \\ \mathbf{x_t} = \mathrm{IE}\left(I_t\right), \end{cases} \tag{2}$$

where IE($\cdot$) and TE($\cdot$) represent the image and text encoders of CLIP, respectively. $\mathbf{x_r}, \mathbf{t_m}, \mathbf{x_t} \in \mathbb{R}^D$ represent the encoded reference image feature, modification text feature, and target image feature, respectively. Then we fuse $\mathbf{x_r}$ and $\mathbf{t_m}$ with an element-wise summation followed by $L2$-normalization as follows,

$$\boldsymbol{\phi} = L2\left(\mathbf{x_r} \oplus \mathbf{t_m}\right), \tag{3}$$

where $\boldsymbol{\phi}$ represents the combined features of $\mathbf{x_r}$ and $\mathbf{t_m}$. $\oplus$ serves as element-wise summation. Finally, to fine-tune the CLIP's image and text encoders, we leverage the widely-used batch-based classification (BBC) loss [2] as follows,

$$L = \frac{1}{B}\sum_{i=1}^{B}\left[-\log\left(\frac{exp\{\kappa(\boldsymbol{\phi}^{(i)}, \mathbf{x_t}^{(i)})/\tau\}}{\sum_{j=1}^{B} exp\{\kappa(\boldsymbol{\phi}^{(j)}, \mathbf{x_t}^{(j)})/\tau}\right)\right], \tag{4}$$

where the subscript i refers to the i-th triplet sample in the mini-batch, B is the batch size, $\kappa\left(\cdot,\cdot\right)$ serves as the cosine similarity function, and τ denotes the temperature factor.

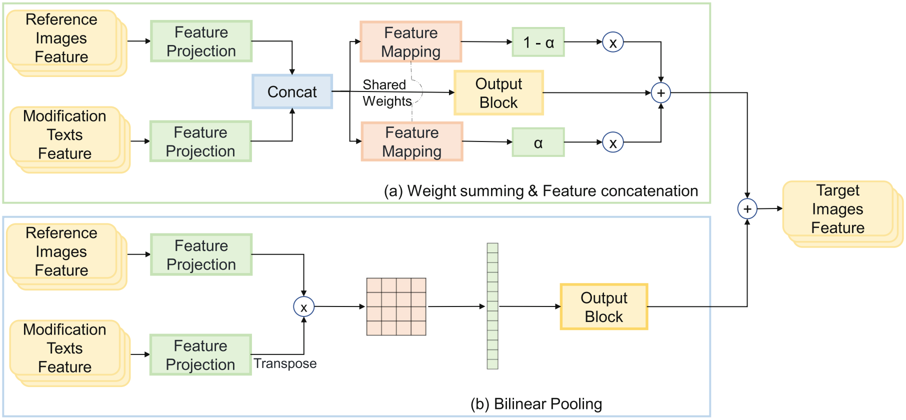

Fig. 2. The architecture of the multimodal fusion module. It takes the reference image feature and the modification texts feature as inputs and outputs a fused representation.

3.2.2 Multimodal Fusion

The overview of the second training stage is depicted in Fig. 1(b). In this stage, we freeze the parameters of CLIP and focus on training a comprehensive multimodal fusion module.

The details of the proposed multimodal fusion module are illustrated in Fig. 2. It employs three strategies for multimodal feature fusion: weighted summing, feature concatenation, and bilinear pooling. Notably, the former two strategies have been explored in Clip4Cir. However, we argue that merely relying on weighted summing and feature concatenation neglects the intricate higher-order interactions inherent in the multimodal query, posing challenges in modeling complex modification intents. To address this limitation, we additionally introduce bilinear pooling as depicted in Fig. 2(b). By leveraging bilinear pooling, the model can comprehensively capture the multimodal query features. This enhancement allows for capturing higher-order interactions between the reference image and modification text features and hence boosts the understanding of the multimodal query.

The bilinear pooling consists of three essential components: the feature projection layer, the bilinear pooling layer, and the output block. First, the features of the reference image and the modification text are processed through the feature projection layers, respectively. Formally, we have,

$$\begin{cases} \mathbf{f}_I = \xi\left(\mathrm{FC}\left(\mathbf{x_r}\right)\right), \\ \mathbf{f}_T = \xi\left(\mathrm{FC}\left(\mathbf{x_m}\right)\right), \end{cases} \tag{5}$$

where $\mathbf{f}_I \in \mathbb{R}^K$ and $\mathbf{f}_T \in \mathbb{R}^K$ refer to the output vectors through the feature projection layer of the reference image and the modification text, respectively. ξ is the RELU activation function and $\mathrm{FC}(\cdot)$ denotes the fully-connected layer. Note that considering the memory cost of storing high dimensional features, we set $K < D$.

In the subsequent bilinear pooling layer, an outer product computation is performed between $\mathbf{f}_I$ and $\mathbf{f}_T$ to obtain a feature map matrix, and this matrix is then flattened into a vector $\mathbf{f}_{bil} \in \mathbb{R}^{K^2}$. It can be formulated as follows,

$$\mathbf{f}_{bil} = \mathrm{Flatten}\left(\mathbf{f}_I \otimes \mathbf{f}_T\right), \tag{6}$$

where $\otimes$ represents the outer product operation. Next, we feed $\mathbf{f}_{bil}$ to the output block, which is as follows,

$$\mathbf{f}_{out} = \mathrm{FC}\left(\xi\left[\mathrm{FC}\left(\mathbf{f}_{bil}\right)\right]\right). \tag{7}$$

Similar to the first stage, we add $\mathbf{f}_{out} \in \mathbb{R}^D$ to the features obtained from the weighted summing and the feature concatenation as the final output of the multimodal fusion module. And the parameters of the multimodal fusion module are optimized by the BBC loss the same as Eq. 4.

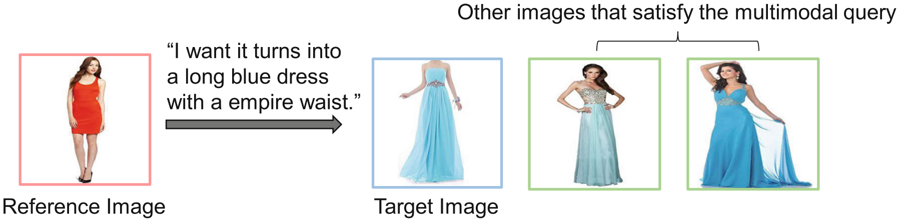
Other images that satisfy the multimodal query
"I want it turns into a long blue dress with a empire waist."
Reference Image
Target Image

Fig. 3. Examples that multiple images satisfy the same multimodal query.

3.2.3 Similarity-Based Data Augmentation

Another major novelty of our work is that we propose a similarity-based data augmentation method, to alleviate the overfitting phenomenon caused by the limited size of training data.

We observe that a single multimodal query could correspond to multiple images that meet the retrieval requirements. Intuitively, these images that align with the same multimodal query tend to exhibit a high similarity degree to the desired target image. As illustrated in Fig. 3, the multimodal query specifies a red dress and intends to modify it into a long blue one with an empire waist. However, apart from the target image, two additional blue dresses meet the multimodal query.

This observation suggests the potential of creating a pseudo triplet by substituting the target image with another similar one. Likewise, replacing the reference image with a similar one can also yield a new pseudo triplet. Specifically, we first utilize the CLIP's image encoder trained from the first stage to extract features of all images in the training dataset, which can be represented as follows,

$$\mathbf{x}_k = \mathrm{IE}\left(I_k\right), k = 1, \cdots, n, \tag{8}$$

where I_k is the image in the training dataset, n denotes the total number of training images, and $\mathbf{x}_k \in \mathbb{R}^D$ is the feature vector corresponding to I_k. Subsequently, we employ the cosine similarity function to calculate the similarity score between the reference/target image and the training images based on the extracted features. To ensure the validity of the pseudo triplets upon image replacement, we design specific constraints from two perspectives: 1) image similarity constraint and 2) triplet matching constraint.

In the image similarity constraint, to ensure the quality of pseudo triplets and distinguish them from the original triplets, we establish the lower similarity threshold ε_l and upper similarity threshold ε_u. Besides, to control the number of pseudo triplets, we follow a method where for each image, only the first s_{max} images with the highest similarity that meet the similarity threshold are selected as its substitute set. The image similarity constraint is as follows,

$$\mathcal{M}_k = \{I_1^{'}, I_2^{'}, ..., I_{s_{max}}^{'}\}, \forall I_i^{'} \in \mathcal{M}_k, \varepsilon_l \leq \kappa(\mathbf{x}_k, \mathbf{x}_i^{'}) \leq \varepsilon_u, \tag{9}$$

where $I_i^{'}$ is a suitable substitute image for I_k, $\mathcal{M}_k$ is the substitute image set for I_k, $\mathbf{x}_i^{'}$ is the feature vector corresponding to $I_i^{'}$. Following this constraint, we can construct the pseudo triplet by replacing the reference image or the target image in the original triplet from their substitute image set. Specifically, the pseudo triplets are denoted as $\tilde{\mathcal{D}}$.

Considering that obtaining similar images directly based on the visual feature outputted by the image encoder could introduce certain noise similar images, we additionally incorporate the triplet matching constraint where we further evaluate the matching degree between the multimodal query and the target image within the pseudo triplets. Specifically, we employ the model derived from the first stage to compute the matching score. If the score surpasses the predefined threshold α, the pseudo triplet is retained; otherwise, it is discarded. Finally, the pseudo triplets are a subset of $\tilde{\mathcal{D}}$, denoted as $\tilde{\mathcal{D}}_{sub}$.

Although we have introduced two constraints to ensure the quality of the pseudo triplets, their quality may be still inferior compared to the original dataset. Therefore, during the second training stage, we adopt an iterative approach utilizing the two types of data. Specifically, we train the model on the original training data $\mathcal{D}$ for k epochs, followed by one training epoch using pseudo data $\tilde{\mathcal{D}}_{sub}$. This training strategy can not only augment the training data to improve the model's generalization ability but also mitigate the adverse effects posed by the lower quality of the pseudo triplets.

4 Experiments

In this section, we first present the experimental settings and then detail the experiments conducted on the Fashion-IQ dataset.

4.1 Datasets and Metrics

Fashion-IQ [24] is a fashion image retrieval dataset based on natural language descriptions. It comprises $77,648$ clothing images and is divided into three subtraining sets: dress, shirt, and top&tee. The training data includes over $18,000$ triplets, where each triplet includes a reference image, a target image, and a modification text. As the test set of Fashion-IQ is not publicly available, we followed the experimental setup of other related works [2,26] in terms of dividing the dataset into training and testing sets.

Following previous efforts [2,23], we adopted the Recall at rank k ($R@k$) as the evaluation metric. It measures the fraction of queries for which the ground truth target is retrieved among the top k results. For all three subsets of the Fashion-IQ dataset, k is set to 10 and 50.

4.2 Implementation Details

Similar to Clip4Cir, we used RN50×4 CLIP as the feature encoder for our model. In the similarity-based data augmentation section, to ensure the validity of the obtained pseudo triplets, we set the lower similarity threshold ε_l and the upper similarity threshold ε_u in Eq. (9) to 90% and 99.5%, respectively. The max number of substitute images s_{max} in Eqn. (9) is set to 4 and the threshold α for the triplet matching constraint is set to 0.6. In the first stage, we employed AdamW optimizer [15] with a learning rate of $2e-6$ and a weight decay coefficient of $1e-2$ to optimize the model. The batch size is set to 64. Additionally, following [18], the temperature factor τ in Eq. (4) is set to 100 to ensure an adequately wide dynamic range of similarity probabilities without interfering with the normal training process. In the second stage, we frozen the CLIP's image and text encoders from the first stage and focused on training the multimodal fusion module. We adopted Adam [11] optimizer with a learning rate of $2e-5$. The batch size is set to $1,024$. All the experiments are implemented by PyTorch, and we fixed the random seeds to ensure reproducibility.

4.3 Performance Comparison

To validate the effectiveness of our method in CIR, we chose the following baselines: TIRG [21], ComAE [1], VAL [3], DATIR [6], CosMo [13], Heteroge [25], SAC [9], DCNet [10], CLVC-Net [23], Clip4Cir [2].

Table 1 summarizes the performance comparison on the Fashion-IQ dataset. From this table, we obtained the following observations. 1) Our proposed method outperforms all baselines over the Fashion-IQ dataset. This confirms the advantages of leveraging a two-stage training approach with a novel fine-tuning strategy and incorporating pseudo triplets. And 2) both our proposed method and Clip4Cir perform better than others without employing visual-language pretraining models, *i.e.*, CLIP. This highlights the crucial role of visual-language pre-training models in this task.

Table 1. Performance comparison on Fashion-IQ. The best results are in boldface, while the second best are underlined.

Method	Dress		Shirt		Top&Tee		Avg	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
TIRG [21]	14.87	34.66	18.26	37.79	19.08	39.62	17.40	37.39
ComAE [1]	14.03	35.10	13.88	34.59	15.80	39.26	19.89	36.31
VAL [3]	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04
DATIR [6]	21.90	43.80	21.90	43.70	27.20	51.60	23.70	46.40
CosMo [13]	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31
Heteroge [25]	26.20	51.20	22.40	46.00	29.30	56.40	26.10	51.20
SAC [9]	26.52	51.01	28.02	51.86	32.70	61.23	29.08	54.70
DCNet [10]	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89
CLVC-Net [23]	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41
Clip4Cir [2]	33.81	59.40	39.99	60.45	41.41	65.37	38.52	61.74
CLIP-CD	**37.68**	**62.62**	**42.44**	**63.74**	**45.33**	**67.72**	**41.82**	**64.79**

Table 2. Ablation study on Fashion-IQ. The results are the average result of the three categories on Fashion-IQ's three sub-training sets.

	Method	R@10	R@50
Stage1	w/o FT	23.44	43.15
	w/ FT-Text	32.71	54.59
	w/ FT-Image	34.02	55.74
	w/ FT-Both	**39.45**	**62.88**
Stage2	w/o Pseudo Data	41.55	64.49
	w/o Bilinear	41.51	64.40
	CLIP-CD	**41.82**	**64.79**

4.4 Ablation Study

To verify the importance of each part of our model, we conducted ablation experiments in two parts: 1) Fine-tune strategy (Stage1) and 2) Component ablation (Stage2).

In the first part, to demonstrate the importance of our fine-tuning strategy, we devised four different fine-tuning experiments which are conducted during the first training stage. In the second part, we compared our proposed method with two other variants of our model to investigate the effectiveness of our key components. Both of them are conducted in the second training stage.

- w/o FT, w/ FT-Text, w/ FT-Image, and w/ FT-Both: To investigate the effectiveness of our fine-tune strategy, we conducted four fine-tuning related variants, including without any CLIP's encoder fine-tuning and with fine-tuning text/ image/ both encoders of CLIP, respectively.
- w/o Pseudo Data: To verify the effectiveness of the proposed similarity-based data augmentation method in the retrieval process, we removed the generated pseudo data and only used the original dataset to train.
- w/o Bilinear: To check the importance of bilinear pooling, we removed bilinear pooling from the multimodal fusion module.

Table 2 shows the ablation results of our proposed method. As can be seen from this table, we gained the following observations. 1) w/ FT-Both achieves better performance than other fine-tune related variants, which confirms the importance of alleviating the domain discrepancy of both text and image simultaneously. 2) Our method surpasses w/o Pseudo Data, indicating that the proposed similarity-based data augmentation method helps improve the model's generalization performance and alleviates the overfitting phenomenon. And 3) w/o Bilinear performs worse compared to our method, which demonstrates that bilinear pooling is useful for boosting the model's multimodal fusion capability.

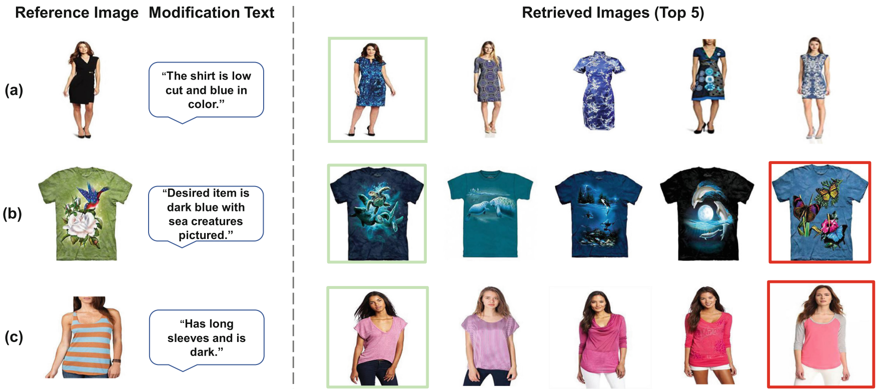
Reference Image
Modification Text
Retrieved Images (Top 5)
(a)
"The shirt is low cut and blue in color."
(b)
"Desired item is dark blue with sea creatures pictured."
(c)
"Has long sleeves and is dark."

Fig. 4. Retrieval examples obtained by our CLIP-CD on Fashion-IQ.

4.5 Case Study

Figure 4 illustrates CIR examples from the three sub-datasets of Fashion-IQ. The top 5 retrieved images are listed, where the green boxes indicate the ground-truth target images labeled in the dataset, whereas the red boxes signify images that fail to meet the retrieval requirements based on our evaluation. It can be observed that the proposed method ranks the ground-truth target image in the first place for all three examples. Specifically, in Fig. 4(a), all the retrieved images can align with the multimodal query. This confirms the effectiveness of our method. Meanwhile, in Fig. 4(b), the 5-th image doesn't exhibit the "sea creatures" trait, even though it shows high similarity to the target image. A similar case can be observed in Fig. 4(c). This may be due to that although we designed two constraints for filtering pseudo samples, a few low-quality triplets might still mislead the optimization of the method. Nevertheless, these observations show that our method succeeds in retrieving the desired target image and most of the retrieved images can meet the multimodal query requirements. This confirms the effectiveness and robustness of our method.

5 Conclusions

In this work, we present a CLIP-based composed image retrieval model with comprehensive fusion and data augmentation, which consists of two training stages. In the first stage, the focus is fine-tuning the image and text encoders of CLIP to alleviate the issue related to domain discrepancy. In the second stage, the emphasis is placed on training a multimodal fusion module to fully integrate the features extracted from the reference image and the modification text. Furthermore, we propose a similarity-based data augmentation method to overcome the problem of insufficient training triplets in the dataset used for this task. Extensive experiments have been conducted on the public Fashion-IQ dataset, and the results demonstrate the effectiveness of our method.

Acknowledgment. The work is supported by the Shandong Provincial Natural Science Foundation (No.:ZR2022YQ59).

References

1. Anwaar, M.U., Labintcev, E., Kleinsteuber, M.: Compositional learning of image-text query for image retrieval. In: Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision, pp. 1140–1149 (2021)
2. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4959–4968 (2022)
3. Chen, Y., Gong, S., Bazzani, L.: Image search with text feedback by visiolinguistic attention learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3001–3011 (2020)
4. Conde, M.V., Turgutlu, K.: Clip-art: contrastive pre-training for fine-grained art classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3956–3960 (2021)
5. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
6. Gu, C., Bu, J., Zhang, Z., Yu, Z., Ma, D., Wang, W.: Image search with text feedback by deep hierarchical attention mutual information maximization. In: Proceedings of the ACM International Conference on Multimedia, pp. 4600–4609 (2021)
7. Guo, J., et al.: HGAN: hierarchical graph alignment network for image-text retrieval. IEEE Trans. Multimedia (2023)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
9. Jandial, S., Badjatiya, P., Chawla, P., Chopra, A., Sarkar, M., Krishnamurthy, B.: SAC: semantic attention composition for text-conditioned image retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4021–4030 (2022)
10. Kim, J., Yu, Y., Kim, H., Kim, G.: Dual compositional learning in interactive image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1771–1779 (2021)
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25 (2012)
13. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 802–812 (2021)
14. Li, X., et al.: Oscar: object-semantics aligned pre-training for vision-language tasks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 121–137. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_8
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

16. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
17. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021)
19. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898 (2014)
20. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Comput. **12**(6), 1247–1283 (2000)
21. Vo, N., et al.: Composing text and image for image retrieval-an empirical odyssey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6439–6448 (2019)
22. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)
23. Wen, H., Song, X., Yang, X., Zhan, Y., Nie, L.: Comprehensive linguistic-visual composition network for image retrieval. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1369–1378 (2021)
24. Wu, H., et al.: Fashion IQ: a new dataset towards retrieving images by natural language feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11307–11317 (2021)
25. Zhang, G., Wei, S., Pang, H., Zhao, Y.: Heterogeneous feature fusion and cross-modal alignment for composed image retrieval. In: Proceedings of the ACM International Conference on Multimedia, pp. 5353–5362 (2021)
26. Zhao, Y., Song, Y., Jin, Q.: Progressive learning for image retrieval with hybrid-modality queries. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1012–1021 (2022)
27. Zhou, Y., Guo, J., Sun, H., Song, B., Yu, F.R.: Attention-guided multi-step fusion: a hierarchical fusion network for multimodal recommendation. arXiv preprint arXiv:2304.11979 (2023)