Cross-modal Feature Alignment and Fusion for Composed Image Retrieval

Yongquan Wan
Shanghai University
yongquanwan@gench.edu.cn

Wenhai Wang
The Chinese University of Hong Kong
wangwenhai362@smail.nju.edu.cn

Guobing Zou
Shanghai University
gbzou@shu.edu.cn

Bofeng Zhang
Shanghai Polytechnic University
bfzhang@sspu.edu.cn

Abstract

Composed Image Retrieval (CIR) presents challenges in expressing search intent through hybrid-modality queries, where users search for a target image using another image along with text to modify certain attributes of the images. CIR encounters two main challenges: cross-modal alignment and feature fusion, due to inherent gaps between images and texts. To address these issues, we decompose the CIR task into a two-stage process and propose the cross-modal feature alignment and fusion model (CAFF). We first fine-tune CLIP's encoders for domain-specific tasks, to learn fine-grained domain knowledge for image retrieval. In the subsequent stage, we enhance the pre-trained model for CIR. Our model incorporates the Image-Guided Global Fusion (IGGF), Text-Guided Global Fusion (TGGF), and Adaptive Combiner (AC) modules. IGGF and TGGF integrate complementary information through intra-modal and inter-modal interactions, discerning alterations in the query image compared to the target image. The AC module balances contributions, yielding the final compositional representation. Extensive experiments on three benchmark datasets demonstrate our model's superiority over state-of-the-art models.

1. Introduction

Interactive image retrieval is emerging as a pivotal technology for search engines, enabling users to express their search intent more efficiently and interactively. Such systems harness the power of algorithms not only to consider the results of previous searches but also to refine the results based on iterative feedback from users, as depicted in Fig. 1. To address this challenge, TIRG [19] first introduced composed image retrieval (CIR), seamlessly merging input images with descriptive text for modifications. This approach effectively utilizes multiple modalities to explicitly capture user search intentions. It is valuable in various contexts, especially in scenarios where users emphasize visual appearance, such as online shopping and fashion retrieval [15].

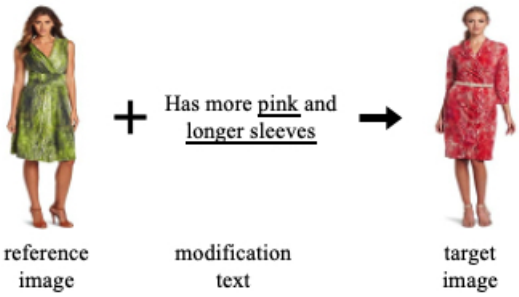

Figure 1. Illustration of a composed image retrieval task.

While CIR can effectively comprehend user search intent, it still faces two primary challenges: (1) cross-modal alignment, which involves aligning images and text with similar semantic in the shared embedding space, and (2) cross-modal fusion, necessitating the editing of image features based on text feedback while retaining original elements. Modification text typically only mentions the parts that need to be changed, implying that the parts not mentioned are intended to be retained in the target image.

Several valuable models [2, 10, 12–14, 20, 22] have been proposed to address these issues. For instance, COSMO [13] introduces a content and style modulator that utilizes decoupled multi-modal non-local blocks to modify image content and infuses global style information into the modified image features. VAL [3] incorporates attention mechanisms across different layers of the deep network to fuse vision and language features. Several studies [1] extend visual language pre-training (VLP) models [17] to tasks such as image retrieval. The typical VLP models are usually trained using large-scale, well-aligned instance-level data from public general domain datasets. However, they may

lack the fine-grained concepts at the attribute-level required for the fashion domain. Therefore, we suggest gradually addressing these challenges, starting by leveraging the CLIP model's capability to learn fine-grained alignment in the fashion domain, and then learning effective multimodal fusion for hybrid-modality queries.

2. Related Work

Composed Image Retrieval. TIRG [19] first proposed the CIR task, utilizing residual gate connections to compose reference image features with text features. Many studies, including COSMO [13], MAAF [5], VAL [3], CLVC [20], SAC [11], and DCNet [12], have explored various approaches to addressing the CIR task. These methods involve techniques such as leveraging multi-level features and integrating multiple sub-networks. Additionally, some studies have investigated the setting of multi-turn of for image retrieval [7, 21]. In this paper, we adopt a single-turn approach to CIR, without loss of generality.

Visual and language pre-training. Vision language pre-training (VLP) models have made remarkable progress in improving the integration of visual and language information. This development offers numerous opportunities and challenges for image retrieval. CIRPLANT [16] is the first to extend VLP model to composed image retrieval tasks in open domains. In the area of fashion image retrieval, several VLP models have emerged [6, 9, 23]. However, these efforts have not been applied to CIR tasks. Baldrati et. al train the CLIP4Cir model [1], a fine-tuned version of the CLIP model [17], for CIR task. Different from CLIP4Cir, we first achieve alignment of image and text concepts at the attribute level by fine-tuning the CLIP encoder, and then learn multimodal composite features.

3. Methods

Given a triplet (I_q, T_m, I_t), where I_q represents the query image, T_m represents the modification text, and I_t represents the target image, we first extract the query image features ϕ_{I_q}, the modification text features ψ_{T_m}, and the target image feature ϕ_{I_t} using the CLIP image and text encoders. We train the model to combine ϕ_{I_q} and ψ_{T_m} to generate the composite feature ϕ_{com}. The goal is to pull the composite feature ϕ_{com} closer to the target feature ϕ_{I_t} in the latent semantic space while pushing away dissimilar images.

Fig. 2 shows the overview of the main architecture of our proposed CAFF model. To progressively learn semantic knowledge from fashion domain data, we propose to train the proposed model in two stages.

3.1. Pre-train Tasks

In stage one, we initialize the image and text encoders using the CLIP backbone and fine-tune the encoders on the fashion domain data with three fashion-specific tasks: fashion image retrieval, fashion attribute prediction, and fashion image category similarity. As shown in Fig. 2 (a), given a fashion dataset $\{I_i, T_i\}$, I_i represents fashion product image, and T_i represents the description text of the corresponding fashion product respectively.

Fashion category similarity enables the model to capture features from both text and image modality by maximizes the similarity between the image and fashion category. We concatenate BERT token [CLS], image category ψ_{cat} and fashion attribute values T_i to form a new text sequence and encode it into $(\psi_{cls}, \psi_{cat}, \psi_{txt})$. We add an adapter (consisting of pooling layer and MLP) after the text encoder and image encoder to learn the adapted features of the category $\widetilde{\psi}^i_{cat}$ and image $\widetilde{\phi}^i_{img}$ respectively, and pull them closer.

Fashion attribute prediction is to label the image with the correct attribute values. Given a fashion dataset $\{I_i, A_i\}$, $A_i = \{a_i^1, ..., a_i^J\}$ denote the attribute value for each attribute type of the image. We attach a attribute prediction head after the image encoder to predict the attributes and treat the attribute prediction task as a multi-label classification task.

Fashion image retrieval aims to measure the similarity for fashion image-text pairs. In image retrieval task, the image feature ϕ_{img} is send to match head to compute the cosine similarity with text feature ψ_{txt}. We use InfoNCE losses to draw similar images and texts closer in the shared embedding space.

3.2. Model Architecture

In stage two, we further train two fusion module and a combiner module to fuse image features and text features for composed image retrieval, as depicted in Fig. 2 (b).

The image-guided global fusion (IGGF) module learn the weights of word in modification text to determine the visual features to retain. The image representation ϕ_{I_q} served as query (Q), and the modification text representation T_m served as key (K) and value (V). The image-guided feature can be calculated through cross-attention layer and feed-forward layer:

$$\psi_{IT}(I_q, T_m) = \phi_{I_q} + MCA(Q, K, V) \,, \tag{1}$$

where MCA denotes the multi-head cross-attention.

The text-guided global fusion (TGGF) aims to learn the visual features relevant to the modification text, which denotes the features to change in the query image. We first project the text feature ψ_{T_m} and the image feature ϕ_{I_q} to obtain the query Q and key K. We concatenate the text and image features and followed with a 1×1 convolution layer to obtain value V. We employ gated dot-product attention to model relationships between features, and add it to ψ_{T_m} to obtain text guided feature ψ_{TI}.

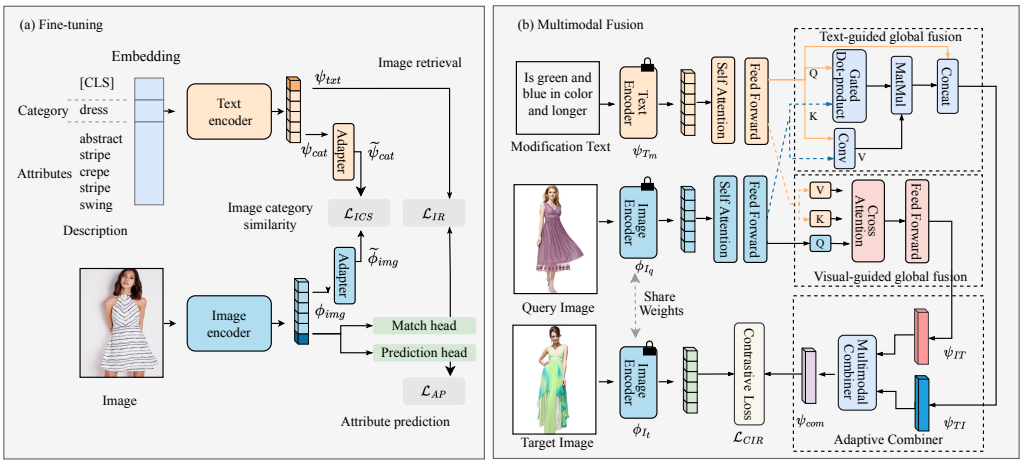

Figure 2. Overview of the proposed model: (a) illustrates our model, which is pre-trained on fashion domain datasets and incorporates three fashion-related tasks; (b) presents our model that is designed specifically for the CIR task. The objective is to learn a composite representation, ϕ_{com}, that aligns with the target image, given a pair of query image, I_q, and modification text, T_m, as input.

$$\psi_{TI}(I_q, T_m) = \psi_{T_m} + softmax(QK^T) \otimes V \,. \quad (2)$$

The adaptive combiner (AC) module generates the final compositional feature ψ_{com} by aggregating the image guided feature and the text guided feature. We follow the similar combiner network in CLIP4Cir [1]. The combiner network is a sum of normalized vectors from three branches. We first project ψ_{TI} and ψ_{IT} through a linear layer and employ a sigmoid function on the outputs of the two branches to learn their convex combination. The concatenation of the two guided features is then sent to the third branch, which has a similar structure. Finally, the final output is obtained as the weighted sum of the outputs from these three branches.

Finally, we train the model by applying end-to-end contrastive learning on the compositional feature ψ_{com} and target image feature ϕ_{I_t}, and optimize the model with InfoNCE loss.

4. Experiments

4.1. Experimental Settings

Datasets. We performed the first stage of pre-training on DeepFashion[1] [15] . To validate the effectiveness of CAFF for CIR against different baseline models, we conducted experiments on three benchmarks, including FashionIQ[2] [21], Fashion200k[3] [8], and Shoes[4] [7].

[1]https://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html
[2]https://github.com/XiaoxiaoGuo/fashion-iq
[3]https://github.com/xthan/fashion-200k.
[4]https://github.com/XiaoxiaoGuo/fashion-retrieval.

Evaluation Metric. Following previous work [3, 12], we used Recall@K (R@K) to evaluate the model performance. The K values are 10 and 50 for FashionIQ, and 1, 10, and 50 for Fashion200k and Shoes.

Table 1. Performance comparison on Fashion200k.

Models	R@1	R@10	R@50
TIRG	14.10	42.50	63.80
VAL	21.20	49.00	68.80
MAAF	18.22	47.52	67.91
COSMO	23.30	50.40	69.30
DCNet	-	46.89	67.56
CLVC	22.60	53.00	72.30
AACL	19.64	58.85	78.86
Artemis	21.50	51.10	70.50
CAFF	**25.21**	**60.17**	**80.79**

Table 2. Performance comparison on Shoes.

Models	R@1	R@10	R@50
TIRG	12.60	45.45	69.39
VAL	16.49	49.12	73.53
MAAF	15.45	49.95	76.36
COSMO	16.72	48.36	75.64
DCNet	-	53.82	79.33
CLVC	17.64	54.39	79.47
Artemis	18.72	53.11	79.31
CAFF	**20.87**	**56.82**	**81.99**

Table 3. Performance comparison of different models on the FashionIQ dataset.

Models	Image encoder	Dress		Shirt		Toptee	
		R@10	R@50	R@10	R@50	R@10	R@50
TIRG [19]	ResNet17	14.87	34.66	18.26	37.89	19.08	39.62
VAL [3]	ResNet50	21.12	42.19	21.03	43.44	25.64	49.49
MAAF [5]	ResNet50	23.80	48.60	21.30	44.20	27.90	53.60
COSMO [13]	ResNet50	25.64	50.30	24.90	48.18	29.21	57.46
DCNet [12]	ResNet50	28.95	56.07	23.95	47.30	30.44	58.29
CLVC [20]	ResNet50×2	29.85	56.47	28.75	54.76	33.50	64.00
AACL [18]	Swin Transformer	29.89	55.85	24.82	48.85	30.88	56.85
Artemis [4]	ResNet50	25.68	51.25	28.59	55.05	21.57	44.13
CLIP4Cir [1]	ResNet50	31.73	56.02	35.77	57.02	36.46	62.77
CAFF	ResNet50	**35.74**	**59.85**	**35.80**	**61.94**	**38.51**	**68.34**

Figure 3. Examples from FashionIQ Shoes, showcasing the top-8 retrieved results. The ground-truth images are indicated by green boxes.

4.2. Performance Comparison

Tab. 1, Tab. 2 and Tab. 3 show the performance on the Fashion200k, Shoes, and FashionIQ datasets, respectively. CAFF demonstrates a significant performance advantage over the competing models. On the FashionIQ dataset, CAFF enhances R@10 and R@50 by 5.86% and 8.16%, respectively. On the Fashion200k dataset, CAFF shows improvements in R@1, R@10, and R@50 by 8.20%, 2.24%, and 2.45%, respectively. On the Shoes dataset, CAFF boosts R@1 by 11.49%, R@10 by 4.47%, and R@50 by 3.17%. CAFF is competitive with the methods that utilize multi-level matching [3, 5], multiple sub-networks [12, 20].

It is worth noting that each model exhibits varying performance across different metrics. For instance, Artemis achieves a higher score in R@1 but lower scores in other metrics on Shoes dataset. This indicates that these models process limited generalization capabilities and struggle to flexibly handle a variety of multimodal queries. In contrast, CAFF demonstrates remarkable versatility, capable of fusing images and modifying text to meet users' retrieval intentions across a broad range of metrics.

4.3. Visualization

Fig. 3 presents quantitative examples from applying our CAFF to the FashionIQ and Shoes datasets. Our model demonstrates the capability to accurately retrieve target images in accordance with textual modifications specified, such as "has long sleeves" or " have different words".

5. Conclusion

To address the challenge of CIR, we propose a novel model named CAFF, designed to align and fuse cross-modal features from image and text modalities. By leveraging the innovative combination of IGGF, TGGF, and AC modules, CAFF enhances its ability to comprehend user search intent. Experimental results demonstrate CAFF's superiority in CIR tasks.

References

[1] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *CVPR*, pages 4959–4968, 2022. 1, 2, 3, 4

[2] Y. Chen and B. Loris. Learning joint visual semantic matching embeddings for language-guided retrieval. In *ECCV*, pages 136–152, 2020. 1

[3] Y. Chen, S. Gong, and Loris B. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, pages 3001–3011, 2020. 1, 2, 3, 4

[4] G. Delmas, R. S. de Rezende, G. Csurka, and D. Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*, 2022. 4

[5] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, and K. Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 2, 4

[6] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, and H. Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *SIGIR*, pages 2251–2260, 2020. 2

[7] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. Feris. Dialog-based interactive image retrieval. In *NIPS*, 2018. 2, 3

[8] X. Han, Z. Wu, P. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, pages 1463–1471, 2017. 3

[9] X. Han, L. Yu, X. Zhu, L. Zhang, Y. Song, and T. Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *ECCV*, pages 634–651, 2022. 2

[10] Y. Hou, E. Vig, M. Donoser, and L. Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *CVPR*, pages 12147–12157, 2021. 1

[11] S. Jandial, P. Badjatiya, P. Chawla, A. Chopra, M. Sarkar, and B. Krishnamurthy. Sac: Semantic attention composition for text-conditioned image retrieval. In *WACV*, pages 4021–4030, 2022. 2

[12] J. Kim, Y. Yu, H. Kim, and G. Kim. Dual compositional learning in interactive image retrieval. In *AAAI*, pages 1771–1779, 2021. 1, 2, 3, 4

[13] S. Lee, D. Kim, and B. Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *CVPR*, pages 802–812, 2021. 1, 2, 4

[14] S. Li, X. Xu, X. Jiang, F. Shen, X. Liu, and H. Shen. Multi-grained attention network with mutual exclusion for composed query-based image retrieval. *TCSVT*, early access:1–15, 2023. 1

[15] Z. Liu, P. Luo, S. Qiu, X.Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016. 1, 3

[16] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould. Image retrieval on real-life images with pretrained vision-and-language models. In *CVPR*, page 2125–2134, 2021. 2

[17] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2

[18] Yuxin Tian, Shawn Newsam, and Kofi Boakye. Fashion image retrieval with text feedback by additive attention compositional learning. In *WACV*, pages 1011–1021, 2023. 4

[19] N. Vo, L. Jiang, C. Sunn, K. Murphy, L. Li, F. Li, and J. Haysm. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, pages 6439–6448, 2019. 1, 2, 4

[20] H. Wen, X. Song, X. Yang, T. Zhan, and L. Nie. Comprehensive linguistic-visual composition network for image retrieval. In *SIGIR*, pages 1369–1378, 2021. 1, 2, 4

[21] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, pages 11307–11317, 2021. 2, 3

[22] Y. Xu, Y. Bin, J. Wei, Y. Yang, G. Wang, and H. Shen. Multi-modal transformer with global-local alignment for composed query image retrieval. *TMM*, 25(1):8346–8357, 2023. 1

[23] M. Zhuge, D. Gao, D. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *CVPR*, pages 12647–12657, 2021. 2