FashionERN: Enhance-and-Refine Network for Composed Fashion Image Retrieval

Yanzhe Chen[1], Huasong Zhong[2], Xiangteng He[1], Yuxin Peng[1]*, Jiahuan Zhou[1], Lele Cheng[2]

[1]Wangxuan Institute of Computer Technology, Peking University
[2]Kuaishou Technology
chenyanzhe@stu.pku.edu.cn, {hexiangteng, pengyuxin, jiahuanzhou}@pku.edu.cn,
{zhonghuasong, chenglele}@kuaishou.com

Abstract

The goal of composed fashion image retrieval is to locate a target image based on a reference image and modified text. Recent methods utilize symmetric encoders (e.g., CLIP) pre-trained on large-scale non-fashion datasets. However, the input for this task exhibits an asymmetric nature, where the reference image contains rich content while the modified text is often brief. Therefore, methods employing symmetric encoders encounter a severe phenomenon: retrieval results dominated by reference images, leading to the oversight of modified text. We propose a Fashion **E**nhance-and-**R**efine **N**etwork (**FashionERN**) centered around two aspects: enhancing the text encoder and refining visual semantics. We introduce a Triple-branch Modifier Enhancement model, which injects relevant information from the reference image and aligns the modified text modality with the target image modality. Furthermore, we propose a Dual-guided Vision Refinement model that retains critical visual information through text-guided refinement and self-guided refinement processes. The combination of these two models significantly mitigates the reference dominance phenomenon, ensuring accurate fulfillment of modifier requirements. Comprehensive experiments demonstrate our approach's state-of-the-art performance on four commonly used datasets.

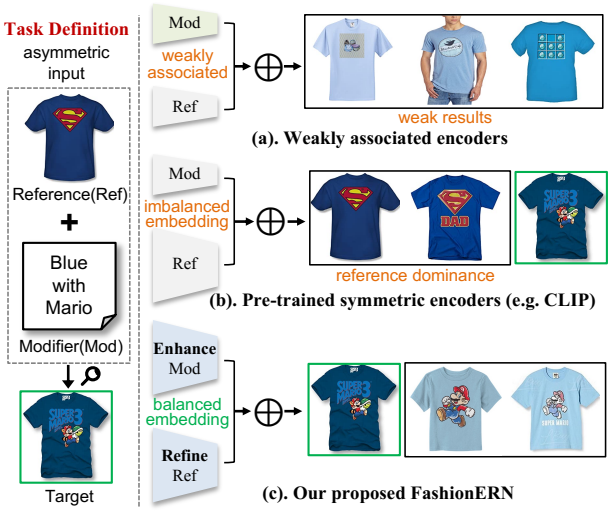


Figure 1: We propose a Fashion Enhance-and-Refine Network (FashionERN) to address the relatively weak results in (a) and reference dominance in (b).

Introduction

Fashion image retrieval has garnered considerable attention as an important e-commerce application (Gajic and Baldrich 2018; Chen et al. 2023a). However, relying solely on fashion images may not meet practical needs, users may modify these images specifically to better match the retrieval target (Chen, Gong, and Bazzani 2020; Guo et al. 2018). To address this, composed fashion image retrieval has emerged as a promising task, jointly retrieving target images with reference images (called reference) and modified text (called modifier) (Kim et al. 2016; Perez et al. 2018; Dodds et al. 2020; Gu et al. 2023), as illustrated in Figure 1.

Existing methods can be categorized into two main categories based on the association between visual and text encoders. The first category involves utilizing two *weakly associated encoders* to represent the textual and visual information, and then combining these representations for target retrieval (Lee, Kim, and Han 2021; Kim et al. 2021; Chen et al. 2022). However, these methods without using pre-trained encoders naturally face the issue of insufficient performance: aligning two modalities with limited fashion data is challenging, especially for image regions or modifiers that are difficult to cover within the limited data. For instance, as illustrated in Figure 1(a), it is challenging to associate "Mario" with the logo region in the image due to limited data. To address the issue above, recent methods (Baldrati et al. 2022; Han et al. 2023a) incorporate *pre-trained symmetric encoders* on larger-scale non-fashion datasets.

Nevertheless, the data in this task significantly differ from the data used for pre-training symmetric encoders. The reference image contains rich contextual information, while the modifier is often more concise than the one used in the pre-training symmetric encoders (Goel et al. 2022). Furthermore, the visual and textual inputs for this task are asymmetric. And the semantics of the modifier are aimed at modifying reference images, which is distinct from the consistent text-visual semantics in symmetric models like

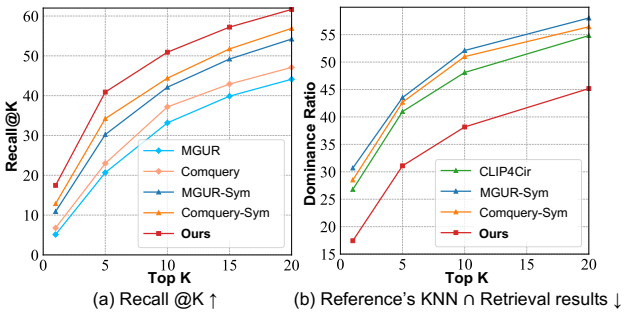
*Corresponding author.

CLIP (Radford et al. 2021). Consequently, the embeddings from reference images and modifiers are unbalanced, leading to sub-optimal retrieval results, where the modifier is easily disregarded and the results are dominated by the reference images. As depicted in Figure 1(b), top-scored retrieval results closely resemble reference images, with the highest-ranked one even being a direct match to the reference. In Figure 2, we visualize the gains from employing pre-trained symmetric encoders and the accompanied Reference Dominance Phenomenon. We select MGUR (Chen et al. 2022) and Comquery (Xu et al. 2023), two representative and open-source methods using weakly associated encoders, and adapt them with symmetric encoders. Figure 2(a) depicts the evident performance improvement from pre-trained symmetric encoders. However, similar to the typical approach of using symmetric encoders like CLIP4Cir (Baldrati et al. 2022), these adapted methods also show the Reference Dominance Phenomenon. We demonstrate the Reference Dominance Phenomenon using the intersection of the reference image's K-nearest neighbors (KNN) and top-K retrieval results in Figure 2(b).

To address the aforementioned challenges and further improve retrieval accuracy, we propose a Fashion Enhance-and-Refine Network (**FashionERN**). Our proposed FashionERN handles the phenomenon of reference image dominance caused by using pre-trained symmetric encoders. This is achieved by dividing the issue into two stages. *In the first stage* we propose a Triple-branch Modifier Enhancement (TME) model which focuses on enriching semantic information and reducing modality discrepancies. Specifically, we introduce a Reference Injection branch to incorporate global content from the reference images into the modifier embedding, such as clothing type, color, size, etc. Additionally, we introduce a Target Alignment branch to map the semantics from the text modality to the visual modality of the target. *The second stage* is to preserve key semantic content from the reference image to achieve a more accurate alignment with the modifier. To address this, we propose a Dual-guided Vision Refinement (DVR) model to progressively refine visual semantics. Specifically, we introduce a Modifier-guided Refine component to select visual regions relevant to the modifier. Additionally, we employ a Self-guided Refine component to filter out non-relevant noise regions in the reference images. Through the aforementioned two stages, we obtain enhanced textual embeddings and refined visual embeddings. The combination of these balanced embeddings results in accurate target retrieval.

The main contributions are summarized as follows:

- We identify and analyze the Reference Dominance Phenomenon caused by adopting pre-trained symmetric encoders in this task, which constrains retrieval performance. Our Fashion Enhance-and-Refine Network (FashionERN) substantially mitigates this issue.
- We propose a Triple-branch Modifier Enhancement (TME) model to enhance the text encoder for enriched information. Moreover, we introduce a Dual-guided Vision Refinement (DVR) model to refine visual semantics for accurate alignment.



(a) Recall @K ↑ (b) Reference's KNN ∩ Retrieval results ↓

Figure 2: (a) Recall Ratio and (b) Dominance Ratio, calculated from the average intersection of reference images' KNN and top-K retrieval results.

- Comprehensive experiments on four commonly used datasets demonstrate our FashionERN as a superior performer compared to previous methods.

Related Work

Composed Fashion Image Retrieval

In recent years, numerous methods have been proposed to solve the challenging composed fashion image retrieval problem. Kim et al. (Kim et al. 2021) paid attention to the little difference between the reference image and the target image, thus modeling the difference between the reference and target image in the embedding space and matched with the embedding of the modifier. Goenka et al. (Goenka et al. 2022) applied VinVL (Zhang et al. 2021) into the proposed method to capture the relationship between the local features of the product and the text. Baldrati et al. (Baldrati et al. 2022) proposed a fusion network to merge the visual and textual features from the CLIP (Radford et al. 2021) network. Xu et al. (Xu et al. 2023) proposed merging information from both local and global dimensions to achieve fine-grained alignment between images. However, the above methods ignore the fact that the modifiers play a guiding role but are usually simple, leading to a biased retrieval outcome in that the results are dominated by the reference image regardless of the crucial semantics in modifiers. In contrast, our approach alleviates the issue by enhancing the text encoder and refining visual semantics.

Multimodal Fusion

Multimodal fusion methods are designed to handle inputs from diverse modalities (Gao et al. 2020; Han et al. 2022b; Chen et al. 2023b) and have been widely applied in downstream tasks such as image captioning (Yang et al. 2023; Brooks, Holynski, and Efros 2023), VQA (Ye, You, and Ma 2022; Dou et al. 2022), etc. We categorize existing visual-textual multimodal fusion methods into two main types. The first type involves fusion between encoders. ALBEF (Li et al. 2021) divides its 12 layers of text encoder into two halves: the first six layers are contrastively learned with the visual encoder, while the latter six layers are utilized as a

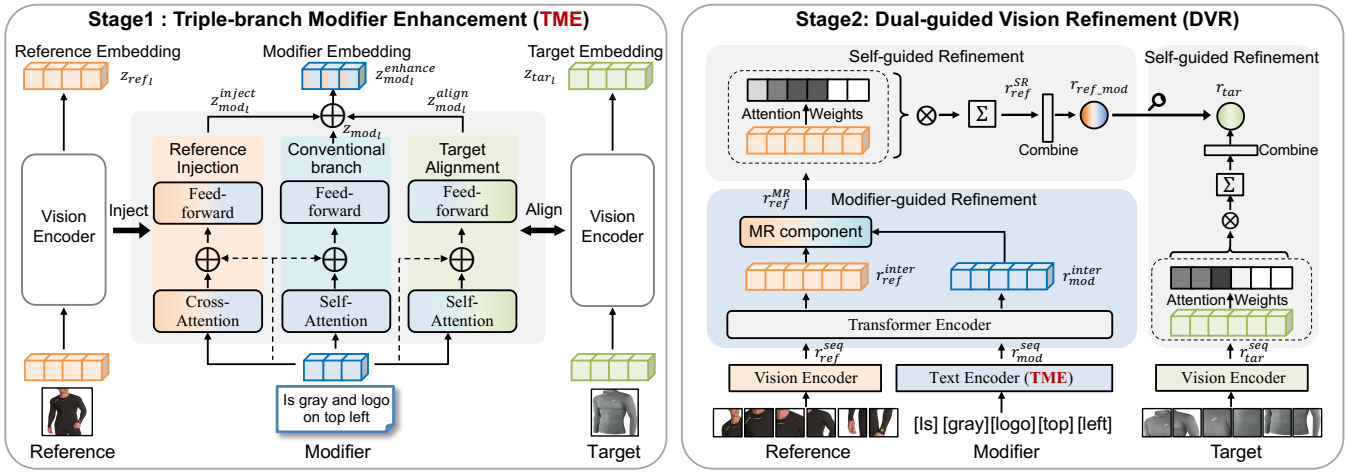


Figure 3: The overall architecture of our proposed FashionERN. The Triple-branch Modifier Enhancement (TME) model employs a tri-branch structure to enhance the semantics of modifiers. The Dual-guided Vision Refinement (DVR) model is designed to refine irrelevant information in reference images.

multimodal encoder for integrating both modalities. Similarly, PMF (Li et al. 2023) employs the last two layers of the Transformer encoder as a multimodal encoder. The second type involves fusion after the encoders. In the case of CLIP4Cir (Baldrati et al. 2022), the visual-textual encoder remains frozen, and an attention-based fusion model is designed to merge features from both modalities. In the task of composed fashion image retrieval, the prevailing methods largely belong to the second category. However, these methods overlook the challenges posed by insufficient textual semantics and excessive visual redundancy, which hinder efficient and precise retrieval. In contrast, our approach attains accurate retrieval through the enrichment of textual semantics while retaining crucial visual semantic content.

Approach

We propose a Fashion Enhance-and-Refine Network (FashionERN) shown in Figure 3, which will be elaborated in the following three sections. In the Triple-branch Modifier Enhancement (TME) section, we present the enhancement of the text encoder through a tri-branch structure. In the Dual-guided Vision Refinement (DVR) section, we elaborate on refining visual semantics using two-step guidance from textual and visual semantics. The Network Training section outlines the process details of training these two models.

Triple-branch Modifier Enhancement (TME)

The TME model consists of three branches: the Conventional branch, the Reference Injection branch, and the Target Alignment branch. These branches are utilized to extract vanilla embeddings, inject reference semantics, and achieve cross-modality alignment with the target, respectively.

Conventional branch: A conventional Transformer encoder comprises an initial embedding layer (word embeddings for textual input and patch embeddings for visual input) followed by a sequence of interleaved layers composed of Multi-Head Self-Attention (MHSA) and MLP blocks. Layer Normalization (LN) is employed preceding each block, and residual connections are established after each block. Both our reference and target encoders utilize a shared encoder. In the modifier encoder, the Conventional branch employs a vanilla encoder structure to extract textual embeddings for modifiers as follows:

$$\begin{cases} z'_{mod_l} = \mathrm{MHSA}(\mathrm{LN}(z_{mod_{l-1}})) + z_{mod_{l-1}} \\ z_{mod_l} = \mathrm{MLP}(\mathrm{LN}(z'_{mod_l})) + z'_{mod_l} \end{cases} \quad (1)$$

where L denotes the number of layers, $l \in [1, L]$.

Reference Injection branch (RI): For the l-th layer encoder, we denote the pre-encoder modifier embedding as $z_{mod_{l-1}}$, and the post-encoder embeddings for reference and target images as z_{ref_l} and z_{tar_l}, respectively. To capture visually relevant information associated with textual semantics, we propose a Reference Injection branch (RI) that utilizes a multi-head cross-attention (CA) block and an MLP block. The branch takes the modifier embedding $z_{mod_{l-1}}$ as the Query and the reference embedding z_{ref_l} as both the Key and Value. This yields an intermediate feature $z^{inject}_{mod_l}$ that incorporates the semantic aspects from the visual embedding, thereby enriching the semantic content of the embedding. The structure of the RI can be depicted as follows:

$$\begin{cases} z^{inject'}_{mod_l} = \mathrm{CA}(z_{mod_{l-1}}, z_{ref_l}, z_{ref_l}) + z_{mod_{l-1}} \\ z^{inject}_{mod_l} = \mathrm{MLP}(\mathrm{LN}(z^{inject'}_{mod_l})) + z^{inject'}_{mod_l} \end{cases} \quad (2)$$

Target Alignment branch (TA): In addition to enhancing the modifier embedding with semantic information from the reference images within the Reference Injection branch, we also enhance it from the perspective of cross-modality alignment. Compared to the reference images, the target images exhibit more semantic consistency with the modifier, yet they belong to different modalities. By introducing KL divergence, we guide the alignment between the text and image modality, aiming to make the semantic distribution of

$z^{align}_{mod_l}$ as close as possible to the target image embeddings z_{tar_l}. The structure of the TA can be depicted as follows:

$$\begin{cases} z^{align'}_{mod_l} = \text{MHSA}(\text{LN}(z_{mod_{l-1}})) + z_{mod_{l-1}} \\ z^{align}_{mod_l} = \text{MLP}(\text{LN}(z^{align'}_{mod_l})) + z^{align'}_{mod_l} \end{cases} \tag{3}$$

The enhanced modifier embedding $z^{enhance}_{mod_l}$ obtained after the l-th layer of the encoder can be represented as follows in our framework:

$$z^{enhance}_{mod_l} = \text{LN}(z_{mod_l} + z^{inject}_{mod_l} + z^{align}_{mod_l}) \tag{4}$$

where z_{mod_l} represents the output of the l-th layer encoder in the Conventional branch. The overall loss function for this stage can be formulated as follows, where $\mathcal{L}_B$ will be introduced in the Network Training section.

$$\begin{aligned} \mathcal{L}_{TME} =& \mathcal{L}_B(z^{enhance}_{mod_L} + z_{ref_L}, z_{tar_L}) + \\ & \sum\nolimits_1^L KL(z^{aign}_{mod_l} || z_{tar_l}) \end{aligned} \tag{5}$$

Dual-guided Vision Refinement (DVR)

The DVR model consists of two components: Modifier-guided Refinement and Self-guided Refinement, aiming at refining visual semantics through two-step guidance from both textual and visual semantics.

Modifier-guided Refinement (MR): In the Dual-guided Vision Refinement model, we maintain the frozen state of the visual encoder and the text encoder enhanced by the TME model. We denote the sequence features extracted by these two encoders as r^{seq}_{ref} and r^{seq}_{mod}, respectively. We begin by concatenating and then passing them through a two-layer Transformer encoder to perform initial interaction between the sequence features of reference images and modifiers. This results in the interacted sequence features r^{inter}_{ref} and r^{inter}_{mod}, shown as follows:

$$[r^{inter}_{ref}; r^{inter}_{mod}] = \text{Transformer}([r^{seq}_{ref}; r^{seq}_{mod}]) \tag{6}$$

where $[r^{inter}_{ref};r^{inter}_{mod}]$ represents the concatenation of features r^{inter}_{ref} and r^{inter}_{mod}. We partition the preliminary interaction features obtained after the Transformer based on the lengths of r^{seq}_{ref} and r^{seq}_{mod}, and denote them as r^{inter}_{ref} and r^{inter}_{mod}, respectively. Next, we propose an MR component that employs a multi-head cross-attention (CA) model together with an MLP model to retain relevant visual features r^{MR}_{ref} based on r^{inter}_{mod}. Taking r^{inter}_{mod} as Query and r^{inter}_{ref} as Key and Value, the process can be illustrated as follows:

$$\begin{cases} r^{MR'}_{ref} = \text{CA}(r^{inter}_{mod}, r^{inter}_{ref}, r^{inter}_{ref}) + r^{inter}_{mod} \\ r^{MR}_{ref} = \text{MLP}(\text{LN}(r^{MR'}_{ref}) + r^{MR'}_{ref}) \end{cases} \tag{7}$$

Self-guided Refinement (SR): After filtering the semantic information of the reference images guided by the modifiers to obtain r^{MR}_{ref}, we further refine it through visual semantic self-guidance. This dual-filtering approach ensures a comprehensive refinement process where the information is successively filtered based on both the textual and visual perspectives, resulting in a more precise alignment with the desired textual and visual semantics. Specifically, we introduce a Self-guided Refinement (SR) component, given the input sequence features r^{MR}_{ref}, we first compute the raw common feature r^{raw}_{ref} by taking the mean across the sequence dimension. Next, a common interaction representation r^{com}_{ref} is obtained using projection functions F_{seq} and F_{com}:

$$r^{com}_{ref} = F_{seq}(r^{MR}_{ref}) \cdot F_{com}(r^{raw}_{ref}) \tag{8}$$

This common interaction representation is then transformed using a shared interaction embedding layer (F_{weight}) followed by a softmax function (σ) to compute attention weights (W):

$$W = \sigma(F_{weight}(r^{com}_{ref})) \tag{9}$$

Finally, the self-refined feature (r^{SR}_{ref}) is computed as a weighted sum of the original sequence features using the computed attention weights:

$$r^{SR}_{ref} = || \sum_{c=1}^{C} W_c \cdot r^{MR}_{ref_c} ||_2 \tag{10}$$

where C represents the sequence length of r^{MR}_{ref}, and W_c denotes the weight of $r^{MR}_{ref_c}$.

Lastly, we propose a Combine module, which adaptively learns weights to combine r^{SR}_{ref} and r^{inter}_{mod}, along with the features z_{ref_L} and $z^{enhance}_{mod_L}$ extracted from the final layer of the vision encoder and the TME text encoder, respectively, resulting in the fused features r_{ref_mod} for both reference images and modifiers:

$$\begin{aligned} r_{ref\_mod} =& \alpha \cdot r^{SR}_{ref} + (1-\alpha) \cdot r^{inter}_{mod} + \\ & \beta \cdot z_{ref_L} + (1-\beta) \cdot z^{enhance}_{mod_L} \end{aligned} \tag{11}$$

where α and β are learnable parameters.

Network Training

In the first step, we initially train the TME model to acquire the enhanced text encoder. The loss function during TME training is depicted in Equation (5). $\mathcal{L}_B$ represents the commonly used batch-based classification loss function in this task (Goenka et al. 2022; Baldrati et al. 2022), where each entry inside a batch acts as a negative sample for all other entries. For a batch of B image-text pairs, the Batch-based Classification loss $\mathcal{L}_B$ is defined as:

$$\mathcal{L}_B(r_{r\_m}, r_t) = \frac{1}{B} \sum_{i=1}^{B} -\log \frac{\exp \kappa(r^i_{r\_m}, r^i_t)}{\sum_{j=1}^{B} \exp \kappa(r^i_{r\_m}, r^j_t)} \tag{12}$$

The kernel κ is the inner product resulting in cosine similarity. During the training of TME, r_t refers to z_{tar_L} extracted from the target images. r_{r_m} is the element-wise summation of extracted features from reference images and modifiers (z_{ref_L} and $z^{enhance}_{mod_L}$), followed by L_2 normalization.

During the second step of training the DVR model, we keep the vision encoder and the enhanced text encoder obtained during the TME stage frozen and only train the MR component, the SR component and the Combine module. In DVR, the loss function is defined as: $\mathcal{L}_{DVR} =$

Method	Dress		Toptee		Shirt		Overall		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Mean
	VAL Evaluation Protocol (ResNet)								
DCNet (Kim et al. 2021)	28.95	56.07	30.44	58.29	23.95	47.30	27.78	53.89	40.84
ARTEMIS (Delmas et al. 2022)	27.16	52.40	29.20	43.64	21.78	54.83	26.05	50.29	38.17
MGUR (Chen et al. 2022)	30.60	57.46	37.37	68.41	31.54	58.29	33.17	61.39	47.28
FashionVLP (Goenka et al. 2022)	32.42	60.29	38.51	68.79	31.89	58.44	34.27	62.51	48.39
Comquery (Xu et al. 2023)	33.86	61.08	42.07	69.30	35.57	62.19	37.17	64.19	50.68
FashionERN (Ours)	**43.93**	**68.77**	**56.09**	**78.38**	**52.70**	**75.07**	**50.91**	**74.07**	**62.49**
	Original Evaluation Protocol (ResNet)								
MGUR (Chen et al. 2022)	24.54	50.12	29.06	55.63	20.70	45.53	24.77	50.43	37.60
FashionVLP (Goenka et al. 2022)	26.77	53.20	28.51	57.47	22.67	46.22	25.98	52.30	39.14
CLIP4Cir (Baldrati et al. 2022)	33.81	59.40	41.41	65.37	39.99	60.45	38.32	61.74	50.03
FashionSAP (Han et al. 2023b)	33.71	60.43	41.91	70.93	33.17	61.33	36.26	64.23	50.25
Css-Net (Zhang et al. 2023)	33.65	63.16	42.65	70.70	35.96	61.96	37.42	65.27	51.35
FashionERN (Ours)	**38.52**	**64.30**	**48.80**	**71.09**	**45.00**	**66.05**	**44.11**	**67.14**	**55.63**
	Original Evaluation Protocol (ViT)								
FashionViL (Han et al. 2022a)	31.53	57.91	36.77	61.81	26.74	50.69	31.68	56.80	44.24
FAME-MiL (Han et al. 2023a)	37.78	63.86	47.22	70.88	45.63	66.78	43.54	67.17	55.36
FashionERN (Ours)	**50.32**	**71.29**	**56.40**	**77.21**	**50.15**	**70.36**	**52.26**	**72.95**	**62.62**

Table 1: Results on FashionIQ dataset. Best scores are highlighted in bold and underlined formats.

$\mathcal{L}_B(r_{ref\_mod}, r_t)$, where r_t encompasses the features of the target image z_{tar_L}, as well as the features resulting from the interaction of the sequence features r_{tar}^{seq} through the SR and the Combine module, represented as: $\gamma \cdot z_{tar_L} + (1-\gamma) \cdot \mathrm{SR}(r_{tar}^{seq})$, where γ is a learnable parameter.

Experiments

We conduct extensive experiments on four commonly used datasets, namely FashionIQ (Yu et al. 2020), Fashion200K (Liu et al. 2021), CIRR (Berg, Berg, and Shih 2010) and Shoes (Han et al. 2017). In the following sections, we present a detailed demonstration of our experimental setup, report the results of our evaluations, and provide comprehensive analyses.

Experimental Setup

Datasets. (1) **FashionIQ** (Yu et al. 2020): The dataset includes 77,684 fashion images categorized into three groups (Dress, Toptee, and Shirt) and organized into triplets. Each triplet comprises a reference image, a target image, and two crowd-sourced captions that explain the variations between the two images. (2) **Fashion200k** (Han et al. 2017): The dataset has 172,000 training and 33,000 testing images. The process used to generate textual feedback involves comparing attributes between image pairs and follows a simple format of "replace [sth] with [sth]." (3) **CIRR** (Liu et al. 2021): The dataset contains 21,552 real-world images from NLVR2 (Suhr et al. 2018). There are 36,554 triplets in total, divided into 3 subsets with 80% in training, 10% in validation, and 10% in testing. (4) **Shoes** (Berg, Berg, and Shih 2010): The dataset has 10,000 training queries and 4,658 validation queries, as split in a previous study (Guo et al. 2018).

Evaluation Metrics. Following the evaluation metrics in (Baldrati et al. 2022; Han et al. 2023a), we adopt the standard top-K recall metric for image retrieval, denoted as R@K. In particular, we use Recall@10 (R@10) and Recall@50 (R@50), along with their average, as the metrics.

Implementation Details. We follow existing work (Baldrati et al. 2022; Han et al. 2023a) and conduct experiments using the same backbones, ResNet50x4 and ViT-B/16. Given that the ResNet model does not inherently have image patch capabilities, we follow KaleidoBERT (Zhuge et al. 2021) to employ a sliding window to obtain the patches. For patch numbers, we apply 2×2 and 3×3 scales to obtain a total of 13 patches. We use Adam (Kingma and Ba 2014) to optimize the network with a mini-batch size of 1024. The initial learning rate is 4e-5, and we adopt a cosine annealing strategy to adjust it. The total number of training epochs is 50. We use 8 Tesla V100 GPUs for model training.

Comparison with State-of-the-art Methods

FashionIQ. To ensure a fair comparison with the previous methods, we adopt two evaluation protocols. The first protocol is referred to as the VAL Evaluation Protocol (Chen, Gong, and Bazzani 2020). In this protocol, all reference and target images are utilized to construct the candidate set. This results in a smaller number of images for retrieval compared to the original validation set, leading to higher performance. The second protocol is the Original Evaluation Protocol proposed in (Yu et al. 2020). As the best-performing method (Han et al. 2023a) under the original evaluation employs the ViT-B/16 model, we also provide a comparative analysis of our approach's results using both backbones.

Table 1 clearly demonstrate that our approach achieves state-of-the-art performance for all three categories in the dataset, surpassing previous methods across all metrics. In particular, for the VAL Evaluation Protocol, our approach improves the average metric by 11.81%. Under the Origi-

Methods	R@10	R@50	Mean
VAL (2020)	49.0	68.8	58.9
DCNet (2021)	46.9	67.6	57.3
CosMo (2021)	50.4	69.3	59.8
FashionVLP (2022)	49.9	70.5	60.2
MGUR (2022)	52.1	70.2	61.2
ARTEMIS (2022)	51.1	70.5	60.8
Css-Net (2023)	50.5	69.7	60.1
Comquery (2023)	52.2	72.2	62.2
FashionERN(Ours)	**54.1**	**72.5**	**63.3**

Table 2: Results on Fashion200k dataset. Best scores are highlighted in bold and underlined formats.

Methods	R@5	S@1	Mean
TIRG (2019)	48.37	22.67	35.52
MAAF (2020)	33.03	21.05	27.04
CIRPLANT (2021)	52.55	39.20	45.88
ARTEMIS (2022)	46.10	39.99	43.05
CompoDiff (2023)	54.36	35.84	45.10
CLIP4Cir (2022)	69.98	68.19	69.59
FashionERN (Ours)	**74.77**	**74.93**	**74.85**

Table 3: Results on CIRR dataset. Best scores are highlighted in bold and underlined formats.

nal Evaluation Protocol setting, when both models employ ResNet as the visual backbone, our approach outperforms the SOTA method Css-Net (Zhang et al. 2023) in terms of the Mean metric by an improvement of 4.28%. Similarly, when both models use ViT-B/16 as the visual backbone, our method achieves a higher Mean score by 7.26% compared to the SOTA method FAME-MiL (Han et al. 2023a). Compared to our baseline CLIP4Cir (Baldrati et al. 2022) and the SOTA methods, our performance enhancement mainly stems from two aspects. Our TME model strengthens the text encoder through two additional branches: injecting information from reference images and reducing modality discrepancies with target images, while the DVR model refines crucial semantics guided by both textual and visual information. Together, they alleviate the dominance of reference-driven results and enable more precise retrieval.

Fashion200k, CIRR and Shoes. The results obtained from the evaluation of our proposed approach on the given dataset are presented in Table 2, Table 3 and Table 4, respectively. Despite the SOTA methods specifically designed for these datasets achieving high performance, our approach still outperforms current methods in all recall measures, which is 1.1%, 5.26% and 1.91% higher than the best available methods in terms of average metrics, respectively. The experimental results on these three datasets further validate the effectiveness and generalizability of our FashionERN.

Ablation Studies

To investigate the effectiveness of our approach, we evaluate the key designs in our model on the FashionIQ dataset. We

Methods	R@10	R@50	Mean
TIRG (2019)	45.45	69.39	57.32
VAL (2020)	49.12	73.53	61.32
CosMo (2021)	48.36	75.64	62.00
FashionVLP (2022)	49.08	77.32	63.20
SAC (2022)	51.73	77.28	64.51
ARTEMIS (2022)	53.11	79.31	66.21
MGUR (2022)	53.63	79.84	66.74
FashionERN (Ours)	**55.59**	**81.71**	**68.65**

Table 4: Results on Shoes dataset. Best scores are highlighted in bold and underlined formats.

Methods	R@10	R@50	Mean
B-R (ResNet)	38.32	61.74	50.03
B-R (stronger encoder)	38.02	63.04	50.53
B-R+TME	41.43	65.02	53.22
B-R+TME+DVR (**Ours**)	44.11	67.14	55.63
B-V (ViT)	44.75	66.78	55.77
B-V (stronger encoder)	45.22	68.01	56.62
B-V+TME	49.28	71.06	60.16
B-V+TME+DVR (**Ours**)	52.26	72.95	62.62

Table 5: Ablation study on FashionIQ dataset of different models."B" implies the baseline method, "-R" implies the use of ResNet and "-V" implies the use of ViT.

conduct incremental experiments with two different backbones (ResNet and ViT) in Table 5. In each group, we add each component incrementally to verify the effect of each component, shown in Table 5. The first row in each group shows the results of the baseline method (Baldrati et al. 2022). For each group, the second row presents the results achieved by employing a stronger text encoder, which is the most straightforward way to enhance the text encoder. Specifically, we substitute the text encoder released by OpenAI with the one trained on LAION (Schuhmann et al. 2021). While this substitution led to certain improvements, it still falls short compared to our TME model due to the absence of semantic enhancement from reference and target images. When DVR is not adopted, r_{r_m} is obtained by element-wise summation of globally extracted features from the reference image and modified text, followed by L_2 normalization. r_t represents the globally extracted features of the target image through the vision encoder, and we use r_{r_m} to retrieve r_t.

Effects of DVR. In Figure 5, we present attention weights' (W) visualization for reference and target images along with their corresponding modifiers. Given that our DVR model preserves the most crucial visual semantics, our approach achieves alignment between keywords in modifiers and corresponding image regions, even when they are relatively small in the whole image. The ability to localize image details is crucial for this task, as modification requirements may involve fine-grained adjustments.

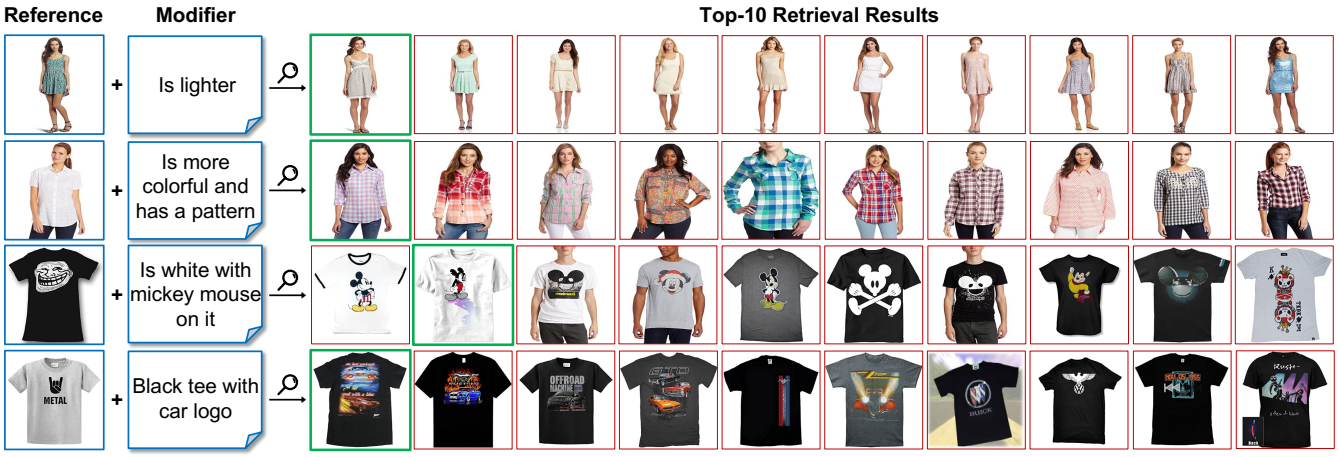


Figure 4: Qualitative results on FashionIQ. We show reference images and modifiers with blue boxes on the left and top-10 retrieval results with descending scores on the right. Ground truths are shown with green boxes, others are shown in red boxes.

Analyses

Qualitative Analyses. We show the references images, related textual feedback and top-10 retrieval results predicted by our approach in Figure 4. *The first row* demonstrates the scenario with particularly simple modifiers. Since our TME model is capable of injecting semantic information from the reference images into the text encoder, e.g., a long floral dress, the color is green, etc., it enables the model to combine the semantics of the modified text and the reference image to accurately retrieve the target. *The second row* shows the case where multiple text conditions are present but each condition description is simple. Our approach not only understands each condition but also takes into account all requirements. *The third row* demonstrates the difficulty of covering the image semantics in the database text. The approach using weakly associated encoders has difficulty in aligning the logo region with the non-occurring requirement of Mickey Mouse. In contrast, our approach preserves the application of pre-trained symmetric encoders while alleviating the reference-dominant retrieval issue, such as retrieving mainly smiley face images. *The last row* shows the fine-grained case where the corresponding modified area is very small. Since our DVR model filters and retains key visual semantic information, our approach is able to accurately align the small area with the car logo requirement.

Analysis of the Reference Dominance Phenomenon. To verify whether the Reference Dominance Phenomenon has been mitigated, we calculate the average proportion of the intersection between the top-K retrieval results and the K-Nearest Neighbor (KNN) of the reference images, as shown in Figure 2(b). We term this average proportion as the Dominance Ratio, which signifies the similarity between retrieval results and reference images. Our FashionERN remarkably decreases the average intersection set proportion by 11.51% compared to the methods adopting pre-trained symmetric encoders. This implies that the phenomenon of retrieval results being dominated by reference images and overlooking modified text has been significantly alleviated.

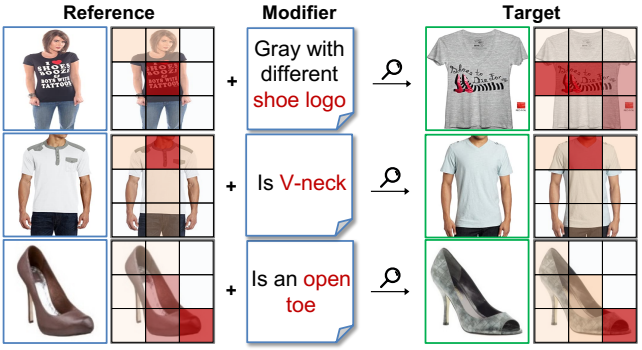


Figure 5: Attention weights' (W) visualization for reference and target images, with their corresponding modifiers.

Conclusion

We present a Fashion Enhance-and-Refine Network (FashionERN) for composed fashion image retrieval. We introduce a Triple-branch Modifier Enhancement model to enrich modified text semantics by injecting reference information and cross-modality alignment. Additionally, our Dual-guided Vision Refinement model preserves key semantic information through two guided refinement processes. Our proposed approach significantly mitigates the reference image dominance phenomenon and effectively fulfills the modifier's requirements, achieving state-of-the-art performance on four datasets. In future work, we will extend our approach to adapt it to fashion image retrieval in multi-round interaction scenarios for better meeting practical needs. Additionally, we will further enhance the user retrieval experience by reducing time complexity.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (61925201, 62132001, U22B2048) and Kuaishou.

References

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21466–21474.

Berg, T. L.; Berg, A. C.; and Shih, J. 2010. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, 663–676. Springer.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Chen, J.; Yuan, H.; Zhang, Y.; He, R.; and Liang, J. 2023a. DCR-Net: Dilated convolutional residual network for fashion image retrieval. *Computer Animation and Virtual Worlds*, 34(2): e2050.

Chen, Y.; Gong, S.; and Bazzani, L. 2020. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3001–3011.

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

Chen, Y.; Zhong, H.; He, X.; Peng, Y.; and Cheng, L. 2023b. Real20M: A Large-scale E-commerce Dataset for Cross-domain Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4939–4948.

Delmas, G.; de Rezende, R. S.; Csurka, G.; and Larlus, D. 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. arXiv:2203.08101.

Dodds, E.; Culpepper, J.; Herdade, S.; Zhang, Y.; and Boakye, K. 2020. Modality-agnostic attention fusion for visual search with text feedback. arXiv:2007.00145.

Dou, Z.-Y.; Kamath, A.; Gan, Z.; Zhang, P.; Wang, J.; Li, L.; Liu, Z.; Liu, C.; LeCun, Y.; Peng, N.; et al. 2022. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35: 32942–32956.

Gajic, B.; and Baldrich, R. 2018. Cross-domain fashion image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1869–1871.

Gao, J.; Li, P.; Chen, Z.; and Zhang, J. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5): 829–864.

Goel, S.; Bansal, H.; Bhatia, S.; Rossi, R.; Vinay, V.; and Grover, A. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35: 6704–6719.

Goenka, S.; Zheng, Z.; Jaiswal, A.; Chada, R.; Wu, Y.; Hedau, V.; and Natarajan, P. 2022. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14105–14115.

Gu, G.; Chun, S.; Kim, W.; Jun, H.; Kang, Y.; and Yun, S. 2023. CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion. arXiv:2303.11916.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris, R. 2018. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31.

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, 1463–1471.

Han, X.; Yu, L.; Zhu, X.; Zhang, L.; Song, Y.-Z.; and Xiang, T. 2022a. Fashionvil: Fashion-focused vision-and-language representation learning. In *European Conference on Computer Vision*, 634–651. Springer.

Han, X.; Zhu, X.; Yu, L.; Zhang, L.; Song, Y.-Z.; and Xiang, T. 2023a. FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2669–2680.

Han, Y.; Zhang, L.; Chen, Q.; Chen, Z.; Li, Z.; Yang, J.; and Cao, Z. 2023b. FashionSAP: Symbols and Attributes Prompt for Fine-grained Fashion Vision-Language Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15028–15038.

Han, Z.; Yang, F.; Huang, J.; Zhang, C.; and Yao, J. 2022b. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20707–20717.

Jandial, S.; Badjatiya, P.; Chawla, P.; Chopra, A.; Sarkar, M.; and Krishnamurthy, B. 2022. SAC: Semantic attention composition for text-conditioned image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4021–4030.

Kim, J.; Yu, Y.; Kim, H.; and Kim, G. 2021. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1771–1779.

Kim, J.-H.; Lee, S.-W.; Kwak, D.; Heo, M.-O.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Multimodal residual learning for visual qa. *Advances in neural information processing systems*, 29.

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

Lee, S.; Kim, D.; and Han, B. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 802–812.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, Y.; Quan, R.; Zhu, L.; and Yang, Y. 2023. Efficient Multimodal Fusion via Interactive Prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2604–2613.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2125–2134.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv:2111.02114.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2018. A corpus for reasoning about natural language grounded in photographs. arXiv:1811.00491.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6439–6448.

Xu, Y.; Bin, Y.; Wei, J.; Yang, Y.; Wang, G.; and Shen, H. T. 2023. Multi-Modal Transformer with Global-Local Alignment for Composed Query Image Retrieval. *IEEE Transactions on Multimedia*.

Yang, Z.; Fang, Y.; Zhu, C.; Pryzant, R.; Chen, D.; Shi, Y.; Xu, Y.; Qian, Y.; Gao, M.; Chen, Y.-L.; et al. 2023. i-code: An integrative and composable multimodal learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10880–10890.

Ye, M.; You, Q.; and Ma, F. 2022. Qualifier: Question-guided self-attentive multimodal fusion network for audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 248–256.

Yu, Y.; Lee, S.; Choi, Y.; and Kim, G. 2020. Curlingnet: Compositional learning between images and text for fashion iq data. arXiv:2003.12299.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

Zhang, X.; Zheng, Z.; Wang, X.; and Yang, Y. 2023. Relieving Triplet Ambiguity: Consensus Network for Language-Guided Image Retrieval. arXiv:2306.02092.

Zhuge, M.; Gao, D.; Fan, D.-P.; Jin, L.; Chen, B.; Zhou, H.; Qiu, M.; and Shao, L. 2021. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12647–12657.