

Exploring Compositional Image Retrieval with Hybrid Compositional Learning and Heuristic Negative Mining

Chao Wang, Ehsan Nezhadarya, Tanmana Sadhu, Shengdong Zhang

Toronto AI Lab

LG Electronics

{chao2.wang, ehsan.nezhadarya, tanmana.sadhu, shengdong.zhang}@lge.com

Abstract

Compositional image retrieval (CIR) is a challenging retrieval task, where the query is composed of a reference image and a modification text, and the target is another image reflecting the modification to the reference image. Due to the great success of the pre-trained vision-and-language model CLIP and its favorable applicability to large-scale retrieval tasks, we propose a CIR model HyCoLe-HNM with CLIP as the backbone. In HyCoLe-HNM, we follow the contrastive pre-training method of CLIP to perform cross-modal representation learning. On this basis, we propose a hybrid compositional learning mechanism, which includes both image compositional learning and text compositional learning. In hybrid compositional learning, we borrow a gated fusion mechanism from a question answering model to perform compositional fusion, and propose a heuristic negative mining method to filter negative samples. Privileged information in the form of image-related texts is utilized in cross-modal representation learning and hybrid compositional learning. Experimental results show that HyCoLe-HNM achieves state-of-the-art performance on three CIR datasets, namely FashionIQ, Fashion200K, and MIT-States.

1 Introduction

In this paper, we explore the task of *compositional image retrieval (CIR)*. As shown in Figure 1, CIR is aimed at retrieving a *target image* slightly different from a *reference image*, where the difference is described by a *modification text*. The key to CIR is to learn the cross-modal composition process from (reference image, modification text) pairs to target images. In the existing CIR models, this is usually

Authors' Contributions: Chao Wang designed and implemented HyCoLe-HNM, and conducted experiments on FashionIQ and Fashion200K; Ehsan Nezhadarya conducted literature review; Tanmana Sadhu conducted experiments on MIT-States; Shengdong Zhang found the limitations of HyCoLe-HNM by conducting experiments on CSS.

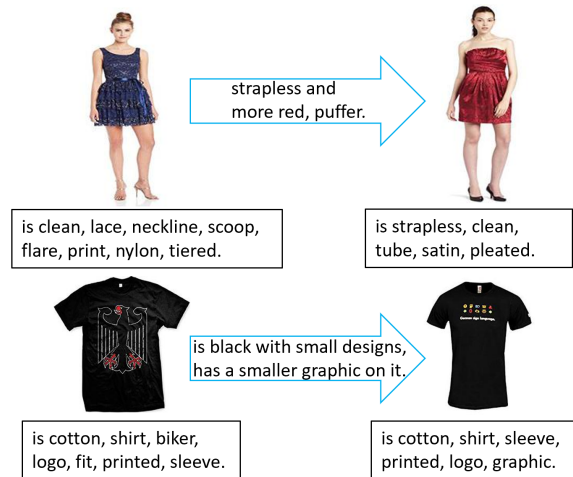


Figure 1: Compositional image retrieval.

realized by fusing and matching image representations obtained from pre-trained vision models with text representations obtained from pre-trained language models (Perez et al., 2018; Vo et al., 2019; Chen and Bazzani, 2020; Dodds et al., 2020; Lee et al., 2021; Kim et al., 2021a; Wen et al., 2021; Anwaar et al., 2021). However, these pre-trained models are pre-trained on uni-modal data, which implies that visual concepts embodied in image representations are not aligned with semantic concepts embodied in text representations. As a result, applying these models to CIR yields limited benefit, which necessitates the application of pre-trained vision-and-language (V&L) models.

CLIP (Radford et al., 2021), a recently-proposed V&L model pre-trained on 400M image-text pairs, has exhibited strong zero-shot performance on image classification. Empirical studies (Kim et al., 2021b; Shen et al., 2022) showed that it is sub-optimal to apply CLIP in a zero-shot manner to complex V&L tasks requiring cross-modal reasoning, such as visual question answering (VQA), visual entailment, and V&L navigation. Meanwhile, Shen et al. (2022) also showed that state-of-the-art (SOTA) performance can be achieved on these

tasks by integrating and fine-tuning CLIP. Due to the requirement for cross-modal compositionality, we believe that CIR is as complex as VQA. Therefore, we use CLIP as the backbone of our proposed CIR model, and fine-tune it together with the rest model components. There are indeed some other pre-trained V&L models than CLIP, such as ALBEF (Li et al., 2021) and BLIP (Li et al., 2022). These models use a single encoder to encode image-text combinations through cross-modal attention mechanisms, which is computationally expensive during retrieval if there are many candidate images. However, CLIP uses two encoders to separately encode images and texts so that we can encode all candidate images in advance and calculate matching scores as simple dot products, which is applicable to large-scale retrieval tasks.

Based on CLIP, we propose a novel CIR model named *HyCoLe-HNM*, which features *hybrid compositional learning* and *heuristic negative mining*. On the one hand, unlike the existing CIR models, which mostly focus on image compositional learning, we propose a hybrid compositional learning mechanism, which includes both image compositional learning and text compositional learning. Specifically, we not only learn the compositional matching between reference images and target images, but also utilize image-related texts as *privileged information* to learn the compositional matching between reference texts and target texts. On the other hand, to facilitate the contrastive optimization of hybrid compositional learning, we also propose a heuristic negative mining method to filter negative samples so that only the negative samples *relevant* to the positive ones are retained. Specifically, we enforce a heuristic rule to identify relevant negative samples, and thereby reduce the space complexity of negative samples from $O(N^3)$ to $O(N^2)$. Compared with hard negative mining methods, the heuristic negative mining method is not only more efficient, but also achieves better performance in the ablation experiments.

For optimal performance, when implementing HyCoLe-HNM, we innovatively integrate some approaches originally aimed at other tasks. Specifically, following the contrastive pre-training method of CLIP, we utilize the above mentioned privileged information to perform *cross-modal representation learning*. Besides, we also borrow a gated fusion mechanism from a question answering (QA) model to perform *compositional fusion*. Experimental re-

sults show that by applying these approaches to HyCoLe-HNM, we achieve SOTA performance on multiple CIR datasets.

2 Model

In this section, we propose our CIR model HyCoLe-HNM. First, we provide a task definition of CIR. Then, we present a cross-modal representation learning method. Next, we propose a hybrid compositional learning mechanism and a heuristic negative mining method, from which HyCoLe-HNM takes its name. Finally, we describe the training and inference of HyCoLe-HNM.

2.1 Task Definition

CIR is to retrieve a target image t from a set of candidate images according to a reference image r and a modification text m , where m describes the change from r to t . We assume that a reference text \tilde{r} and a target text \tilde{t} , which separately embody the semantics of r and t , are provided as privileged information for training. For example, \tilde{r} and \tilde{t} can be image-related captions. However, such information is not available for inference.

2.2 Cross-Modal Representation Learning

To map images and texts into a joint representation space, we utilize privileged information in the form of image-related texts to jointly train an image encoder and a text encoder through cross-modal representation learning. As shown in Figure 2a, we use an image encoder to encode reference images and target images, and use a text encoder to encode reference texts and target texts. To benefit from V&L pre-training, we use CLIP as the backbone of both encoders. Specifically, we use the image part of CLIP, which is a vision transformer (ViT) (Dosovitskiy et al., 2020), as the backbone of the image encoder, and use the text part of CLIP, which is a GPT-like (Radford et al., 2018) language model, as the backbone of the text encoder. Besides, we also add a linear projection layer after the backbone of each encoder, and apply L2 normalization to the output of each linear projection layer. The output dimensionality of both linear projection layers is d , which is the dimensionality of the joint representation space.

To learn the joint representation space, we adopt the InfoNCE loss (Oord et al., 2018) used in the contrastive pre-training of CLIP, and apply it to both the reference side and the target side. Specif-

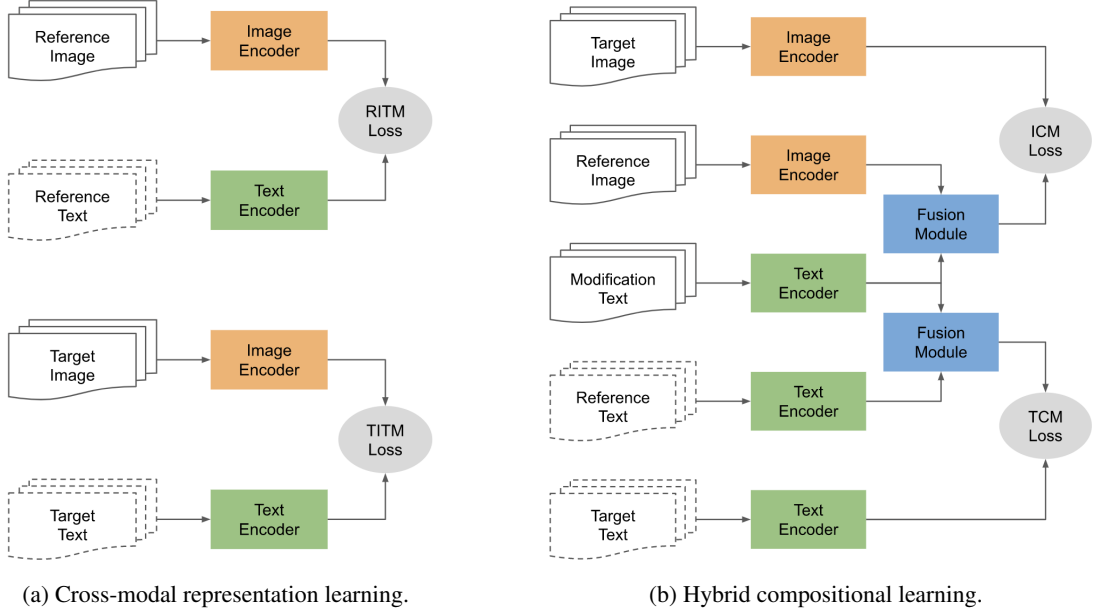


Figure 2: Our proposed HyCoLe-HNM model. Components with the same color share parameters.

ically, given a mini-batch of N reference image-text pairs $\{(r_1, \tilde{r}_1), \dots, (r_N, \tilde{r}_N)\}$, we treat them as positive samples, and generate $N^2 - N$ negative samples by replacing the text \tilde{r}_i in each positive sample (r_i, \tilde{r}_i) separately with the other $N - 1$ texts $\{\tilde{r}_1, \dots, \tilde{r}_N\} - \{\tilde{r}_i\}$. For each of the positive samples and negative samples, we calculate the cosine similarity between the image representation and the text representation, and thereby construct a reference image-text matching (RITM) similarity matrix $S_{RITM} \in \mathbb{R}^{N \times N}$, where the element at the i -th row and the j -th column corresponds to the sample (r_i, \tilde{r}_j) . Analogously, given a mini-batch of N target image-text pairs $\{(t_1, \tilde{t}_1), \dots, (t_N, \tilde{t}_N)\}$, we construct a target image-text matching (TITM) similarity matrix $S_{TITM} \in \mathbb{R}^{N \times N}$. Obviously, the diagonal elements in the two matrices correspond to the positive samples, while the off-diagonal elements correspond to the negative samples. On this basis, we minimize the following RITM loss \mathcal{L}_{RITM} and TITM loss \mathcal{L}_{TITM} so that in the learned joint representation space, an image and a text are close to each other if they are paired, and apart from each other if not:

$$\mathcal{L}_{RITM} = \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{RITM}}{\tau_{RITM}} \right) \right) \right) + \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{RITM}^\top}{\tau_{RITM}} \right) \right) \right) \quad (1)$$

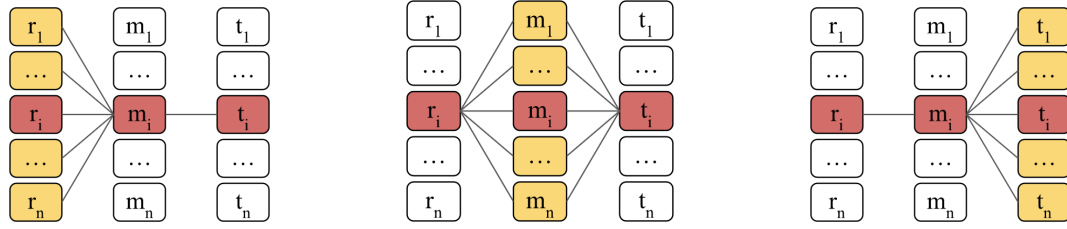
$$\mathcal{L}_{TITM} = \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{TITM}}{\tau_{TITM}} \right) \right) \right) + \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{TITM}^\top}{\tau_{TITM}} \right) \right) \right) \quad (2)$$

where τ_{RITM} and τ_{TITM} are trainable temperatures, $\text{tr}(\cdot)$ denotes calculating matrix trace, and $\text{softmax}(\cdot)$ is calculated along each row.

2.3 Hybrid Compositional Learning and Heuristic Negative Mining

The existing CIR models mostly focus on image compositional learning, which is to learn the compositional matching between reference images and target images conditioned on modification texts. In our proposed CIR model, besides image compositional learning, we also utilize privileged information in the form of image-related texts to perform text compositional learning, which is to analogously learn the compositional matching between reference texts and target texts, and thus name this mechanism hybrid compositional learning. As shown in Figure 2b, based on cross-modal representation learning, we use a fusion module to fuse modification text representations separately into reference image representations and reference text representations, which can be seen as compositional fusion. To implement the fusion module, we borrow the following gated fusion mechanism from Wang et al. (2018), which was originally proposed to address the task of QA, and apply L2 normalization to its output:

$$\begin{aligned} f(x, y) &= \text{norm}(g \odot h + (1 - g) \odot x) \\ g &= \text{sigmoid}(W_g[x; y; x \odot y; x - y] + b_g) \\ h &= \text{gelu}(W_h[x; y; x \odot y; x - y] + b_h) \end{aligned}$$



(a) Reference-based negative mining. (b) Modification-based negative mining. (c) Target-based negative mining.

Figure 3: Our proposed heuristic negative mining method.

where W_g and W_h are trainable weight matrices, b_g and b_h are trainable bias vectors, $f(x, y)$ denotes fusing y into x , \odot denotes element-wise multiplication, $\text{norm}(\cdot)$ denotes L2 normalization, and $[\cdot]$ denotes vector concatenation.

As in cross-modal representation learning, we also adopt the InfoNCE loss in hybrid compositional learning. Specifically, in image compositional learning, given a mini-batch of N (reference image, modification text, target image) triples $\{(r_1, m_1, t_1), \dots, (r_N, m_N, t_N)\}$, we treat them as positive samples, and generate $N^3 - N$ negative samples by enumerating the other possible triples $\{r_1, \dots, r_N\} \times \{m_1, \dots, m_N\} \times \{t_1, \dots, t_N\} - \{(r_1, m_1, t_1), \dots, (r_N, m_N, t_N)\}$. However, most of these negative samples are easy negatives, which are irrelevant to the positive samples and thus have little effect on the contrastive optimization. Therefore, we filter these negative samples to only retain the hard negatives, which are relevant to the positive samples. Instead of applying hard negative mining methods, we propose a more efficient heuristic negative mining method, which is to identify relevant negative samples by enforcing a heuristic rule: **a negative sample is relevant if and only if it is different from a positive sample in either the reference image, the modification text, or the target image**. As shown in Figure 3, we implement this rule as the following three operations, which reduce the space complexity of negative samples from $O(N^3)$ to $O(N^2)$:

- **Reference-based negative mining.** For each positive sample, we select the $N - 1$ negative samples that only differ in the reference image. This operation yields $N^2 - N$ relevant negative samples in total.
- **Modification-based negative mining.** For each positive sample, we select the $N - 1$ negative samples that only differ in the modification text. This operation yields $N^2 - N$

relevant negative samples in total.

- **Target-based negative mining.** For each positive sample, we select the $N - 1$ negative samples that only differ in the target image. This operation yields $N^2 - N$ relevant negative samples in total.

For each of the positive samples and relevant negative samples, we first fuse the modification text representation into the reference image representation, and then calculate the cosine similarity between the fusion result and the target image representation. In this way, for the above three operations, we construct three image compositional matching (ICM) similarity matrices $S_{R-ICM} \in \mathbb{R}^{N \times N}$, $S_{M-ICM} \in \mathbb{R}^{N \times N}$, and $S_{T-ICM} \in \mathbb{R}^{N \times N}$, where the elements at the i -th row and the j -th column separately corresponds to the samples (r_j, m_i, t_i) , (r_i, m_j, t_i) , and (r_i, m_i, t_j) . Obviously, the diagonal elements in the three matrices correspond to the positive samples, while the off-diagonal elements correspond to the relevant negative samples. On this basis, we minimize the following ICM loss \mathcal{L}_{ICM} so that the compositional matching between a reference image and a target image conditioned on a modification text is promoted if the modification text reflects the change from the reference image to the target image, and suppressed if not:

$$\begin{aligned} \mathcal{L}_{ICM} = & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{R-ICM}}{\tau_{ICM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{R-ICM}^\top}{\tau_{ICM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{M-ICM}}{\tau_{ICM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{M-ICM}^\top}{\tau_{ICM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{T-ICM}}{\tau_{ICM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{T-ICM}^\top}{\tau_{ICM}} \right) \right) \right) \end{aligned} \quad (3)$$

where τ_{ICM} is a trainable temperature.

Analogously, in text compositional learning, given a mini-batch of N (reference text, modification text, target text) triples $\{(\tilde{r}_1, m_1, \tilde{t}_1), \dots, (\tilde{r}_N, m_N, \tilde{t}_N)\}$, we apply the same method as in image compositional learning to construct three text compositional matching (TCM) similarity matrices $S_{R-TCM} \in \mathbb{R}^{N \times N}$, $S_{M-TCM} \in \mathbb{R}^{N \times N}$, and $S_{T-TCM} \in \mathbb{R}^{N \times N}$, and thereby minimize the following TCM loss \mathcal{L}_{TCM} :

$$\begin{aligned} \mathcal{L}_{TCM} = & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{R-TCM}}{\tau_{TCM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{R-TCM}^\top}{\tau_{TCM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{M-TCM}}{\tau_{TCM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{M-TCM}^\top}{\tau_{TCM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{T-TCM}}{\tau_{TCM}} \right) \right) \right) + \\ & \frac{1}{N} \text{tr} \left(-\log \left(\text{softmax} \left(\frac{S_{T-TCM}^\top}{\tau_{TCM}} \right) \right) \right) \end{aligned} \quad (4)$$

where τ_{TCM} is a trainable temperature.

Since our proposed CIR model features hybrid compositional learning and heuristic negative mining, we name it HyCoLe-HNM.

2.4 Training and Inference

To train HyCoLe-HNM, we minimize the following total loss \mathcal{L} through gradient descent:

$$\mathcal{L} = \mathcal{L}_{ICM} + \alpha \mathcal{L}_{TCM} + \beta (\mathcal{L}_{RITM} + \mathcal{L}_{TITM}) \quad (5)$$

where the loss scaling factors α and β are hyper-parameters. To effectively fine-tune CLIP, we set the backbone learning rate as the product of the global learning rate and a backbone activity ratio γ , which is another hyper-parameter. From a knowledge perspective, γ controls the trade-off between the knowledge transferred from CLIP and that embodied in the training data. For inference, we encode all candidate images in advance and cache the resulting representations. On this basis, for each given (reference image, modification text) pair, we perform compositional fusion, calculate the cosine similarity between the fusion result and each candidate image representation as the matching score, and thereby rank all candidate images according to the resulting matching scores.

3 Related Works

3.1 Image Retrieval

Given a query, image retrieval methods can retrieve the most similar images to the query, from an image database. In real-life scenarios, users may use

different types of queries to search for an image. Conventional image retrieval methods are based on the assumption that the input query is of a single type or modality. Some examples include queries of type image (Dubey, 2021; Liu et al., 2016), text (Tan et al., 2019; Lu et al., 2019; Messina et al., 2021; Wang et al., 2016), attribute (Zhao et al., 2017) and sketch (Sangkloy et al., 2016; Radenovic et al., 2018; Sain et al., 2021).

3.2 Compositional Learning

The main idea behind compositional learning is to develop a complex concept by combining multiple primitive concepts (Misra et al., 2017). Compositional learning is widely explored in different cross-modal tasks, such as image captioning (Zhou et al., 2020; Zhang et al., 2021) and VQA (Antol et al., 2015; Zhou et al., 2021).

Recently, CIR has gained a lot of attention more specifically for fashion product search (Wu et al., 2021). Augmenting an image query with additional modification text input for image retrieval has been the main line of work in this area. TIRG (Vo et al., 2019) applies compositional learning to image retrieval, using a residual gating mechanism to fuse image and text representations. To compose the vision and language content, VAL (Gu et al., 2021) plugs composite transformers into convolution layers at different depths of the network. MAAF (Dodds et al., 2020) concatenates image and text tokens and passes them into a Transformer encoder-like architecture. Hosseinzadeh and Wang (2020) apply self-attention to image and text representations independently and use cross-attention fusion between the two representations. To change the image content and style based on the modification text, CoSMo (Lee et al., 2021) applies content modulator (CM) and style modulator (SM) to the reference image.

JVSM (Chen and Bazzani, 2020) jointly learns image-text representations as well as compositional representations in a unified embedding space using a multi-task learning framework. Similar to our method, privileged information is used at training time. However, unlike our method which is based on both cross-modal (image-text) and uni-modal (text-text) compositional learning, they only use cross-modal compositionality at training time. Although using the cross-modal compositional learning plays the main role in the performance of the proposed method, we show that language composi-

tionality further improves the results.

3.3 Vision-and-Language Pre-Training

The recent success of Transformer-based language model pre-training (Lan et al., 2019; Clark et al., 2019) has inspired vision-and-language (V&L) pre-training in different tasks, such as VQA, image captioning, visual commonsense reasoning and image retrieval (Chen et al., 2020b; Sun et al., 2021; Li et al., 2020; Radford et al., 2021). The main objective of V&L pre-training is to construct a cross-modal representation space to help improve the generalizability and sample efficiency of downstream tasks by training on large-scale image-text datasets.

V&L pre-training has also been applied to CIR. CIRPLANT (Liu et al., 2021) uses the pre-trained V&L model OSCAR (Li et al., 2020) as the composition module. The method achieves SOTA performance on the authors’ created CIR dataset. However, its performance on FashionIQ (Wu et al., 2021) is sub-optimal, apparently due to the domain shift between the pre-training dataset and FashionIQ.

Recently, CLIP (Radford et al., 2021) has been proposed to learn visual concepts with language supervision. It follows a late fusion design where image and text representations, encoded by independent image and text encoders, are learned using a contrastive loss. Due to the success of CLIP in different V&L tasks (Shen et al., 2022), we employ pre-trained CLIP as a backbone model for the proposed method. Experimental results show that the proposed CLIP-based text-guided image retrieval method achieves SOTA performance on different datasets.

3.4 Learning with Privileged Information

Privileged information refers to the information which is available at training time but not at test time. The paradigm of learning with privileged information was first formulated by Vapnik and Vashist (2009). The privileged information is used in different tasks such as object detection (Hoffman et al., 2016; Mordan et al., 2018), semantic segmentation (Lee et al., 2018) and image super-resolution (Lee et al., 2020) to train a stronger model. Recently, side information in the form of attributes or image caption has been used to improve the performance of image retrieval methods (Wu et al., 2021; Chen and Bazzani, 2020). Similar to these

methods, we use the attributes provided for each image as privileged information.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

To verify the effectiveness of HyCoLe-HNM, we conduct experiments on three CIR datasets, namely FashionIQ (Wu et al., 2021), Fashion200K (Han et al., 2017), and MIT-States (Isola et al., 2015). We pre-process these datasets into a unified format, where each data sample consists of a reference image, a reference text, a target image, a target text, and a modification text (refer to Appendix A for the data statistics and examples of each dataset). Besides, we also adopt recall-at-K (R@K) as unified evaluation metrics on these datasets, which is the percentage of data samples whose target image appears in the top-K retrieved images.

4.1.2 Implementation Details

We use PyTorch (Paszke et al., 2019) to implement HyCoLe-HNM, use Ray’s Tune (Liaw et al., 2018) to perform hyper-parameter optimization, and use HuggingFace’s Transformers (Wolf et al., 2019) to load CLIP. We construct HyCoLe-HNM separately with three versions of CLIP, namely CLIP-ViT-B/32, CLIP-ViT-B/16, and CLIP-ViT-L/14. For optimization, we apply an AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of 0.0001, a weight decay factor of 0.01, and a mini-batch size of 64. The trainable temperatures τ_{RITM} , τ_{TITM} , τ_{ICM} , and τ_{TCM} are initialized to e^{-1} , the loss scaling factors α and β are separately set to 0.4 and 0.1, and the backbone activity ratio γ is set to 0.001. We optimize the model for 64 epochs on a single NVIDIA V100 GPU, where a cosine schedule is used to anneal the learning rate after 6 warm-up epochs. Besides, to improve the efficiency, we also apply mixed precision training and gradient checkpointing. For evaluation, we follow Vo et al. (2019) to group candidate images, and thereby treat candidate images in the same group as identical.

4.2 Experimental Results

4.2.1 FashionIQ

FashionIQ is a dataset of fashion images, which fall into three categories, namely dresses, shirts, and tops&tees. This dataset is organized as (reference image, target image) pairs, where each pair

Model	Dresses		Shirts		Tops&Tees		Average		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	(R@10+R@50)/2
JVSM	10.7	25.9	12.0	27.1	13.0	26.9	11.9	26.63	19.27
FILM	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04	25.28
Relationship	15.44	38.08	18.33	38.63	21.10	44.77	18.29	40.49	29.39
TIRG	20.02	44.55	25.73	49.88	26.72	54.82	24.16	49.75	36.96
VAL	22.53	44.00	22.38	44.15	27.53	51.68	24.15	46.61	35.38
CoSMo	25.64	50.30	24.90	49.18	29.21	57.46	26.53	52.31	39.42
CIRPLANT	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
MAAF	23.8	48.6	21.3	44.2	27.9	53.6	24.33	48.8	36.57
HyCoLe-HNM (CLIP-32)	38.25	64.04	41.17	66.29	48.75	74.86	42.72	68.39	55.55
HyCoLe-HNM (CLIP-16)	39.34	65.38	45.00	69.09	50.13	76.64	44.82	70.37	57.60
HyCoLe-HNM (CLIP-14)	41.38	68.92	46.52	71.34	54.67	77.41	47.52	72.55	60.03

Table 1: Performance comparison on FashionIQ with JVSM (Chen and Bazzani, 2020), FiLM (Perez et al., 2018), Relationship (Santoro et al., 2017), TIRG (Vo et al., 2019), VAL (Chen et al., 2020a), CosMo (Lee et al., 2021), CIRPLANT (Liu et al., 2021) and MAAF (Dodds et al., 2020).

comes with two human-written relative captions, and each image comes with an attribute set. For each pair, we denote the relative captions by z_1 and z_2 , denote the attribute set of the reference image by $\{u_1, \dots, u_p\}$, and denote that of the target image by $\{v_1, \dots, v_q\}$. We generate “ z_1, z_2 .” as the modification text, generate “ $is u_1, \dots, u_p$.” as the reference text, and generate “ $is v_1, \dots, v_q$.” as the target text. We optimize HyCoLe-HNM on all training samples, and evaluate it on the test samples of each category. As shown in Table 1, HyCoLe-HNM outperforms the existing CIR methods by a large margin. Some retrieval examples are shown in Figure 4a.

4.2.2 Fashion200K

Model	R@1	R@10	R@50
Han et al.	6.3	19.9	38.3
Show and Tell	12.3	40.2	61.8
Param Hashing	12.2	40.0	61.7
Relationship	13.0	40.5	62.4
FiLM	12.9	39.5	61.9
TIRG	14.1	42.5	63.8
TIRG+BERT	19.9	51.7	71.8
ComposeAE	22.8	55.3	73.4
JVSM	19.0	52.1	70.0
VAL	22.9	50.8	72.7
CoSMo	23.3	50.4	69.3
HyCoLe-HNM (CLIP-32)	22.1	66.9	87.5
HyCoLe-HNM (CLIP-16)	23.5	69.7	90.4
HyCoLe-HNM (CLIP-14)	26.2	72.4	91.3

Table 2: Performance comparison on Fashion200K with Han et al. (2017), Show and Tell (Vinyals et al., 2015), Param Hashing (Noh et al., 2016), Relationship, FiLM, TIRG, TIRG+BERT (Anwaar et al., 2021), ComposeAE, JVSM, VAL, and CoSMo.

Fashion200K is another dataset of fashion images,

which fall into five categories, namely pants, skirts, dresses, tops, and jackets. Similar to FashionIQ, each image in this dataset comes with an attribute set. Following Vo et al. (2019), we traverse all possible image pairs in each category to select (reference image, target image) pairs. Specifically, we select an image pair (i_1, i_2) if the attribute set of i_1 differs from that of i_2 in only one attribute. In this case, we denote the different attribute of i_1 by u , and denote that of i_2 by v . We generate “ $is not u, is v$.” as the modification text, and generate the reference text and the target text in the same way as in FashionIQ. We optimize HyCoLe-HNM on all training samples, and evaluate it on the test samples provided by Vo et al. (2019). As shown in Table 2, for R@10 and R@50, HyCoLe-HNM outperforms the existing CIR models by a large margin. For R@1, HyCoLe-HNM is comparable with the SOTA CIR models when using the base CLIPs (CLIP-ViT-B/32 and CLIP-ViT-B/16), but much better when using the large CLIP (CLIP-ViT-L/14). Some retrieval examples are shown in Figure 4b.

4.2.3 MIT-States

MIT-States is a dataset of object images, where each image comes with a noun specifying the object name and an adjective describing the object state. Following Vo et al. (2019), we traverse all possible image pairs to select (reference image, target image) pairs. Specifically, we select an image pair (i_1, i_2) if i_1 and i_2 have the same noun but different adjectives. In this case, we denote the noun by o , denote the adjective of i_1 by u , and denote that of i_2 by v . We generate “ $is not u, is v$.” as the modification text, generate “ $u o$.” as the reference text, and generate “ $v o$.” as the target

Model	R@1	R@5	R@10
Show and Tell	11.9	31.0	42.0
Att as Operator	8.8	27.3	39.1
Relationship	12.3	31.9	42.9
FiLM	10.1	27.7	38.3
TIRG	12.2	31.9	43.1
TIRG+BERT	13.3	34.5	46.8
ComposeAE	13.9	35.3	47.9
MAAF	12.7	32.6	44.8
Locally Bounded	14.7	35.3	46.6
HyCoLe-HNM (CLIP-32)	16.1	38.5	50.9
HyCoLe-HNM (CLIP-16)	17.8	42.2	54.6
HyCoLe-HNM (CLIP-14)	19.5	44.4	56.8

Table 3: Performance comparison on MIT-States with Show and Tell, Att as Operator (Nagarajan and Grauman, 2018), Relationship, FiLM, TIRG, TIRG+BERT, ComposeAE, MAAF, and Locally Bounded (Hossein-zadeh and Wang, 2020).

text. With the data splitting provided by Vo et al. (2019), we optimize HyCoLe-HNM on all training samples, and evaluate it on all test samples. As shown in Table 3, HyCoLe-HNM outperforms the existing CIR models, where the advantage is more significant when using the large CLIP than when using the base CLIPs. Some retrieval examples are shown in Figure 4c.

4.3 Ablation Study

Model	Overall Performance		
	FashionIQ	Fashion200K	MIT-States
HyCoLe-HNM	52.16	59.55	35.19
w/o Text Compositional Learning	51.53	59.29	35.02
w/o Heuristic Negative Mining	36.42	52.55	34.58
w/o Gated Fusion	41.72	55.58	32.97
w/o Privileged Information	51.25	58.71	34.53
Frozen CLIP	47.29	51.93	31.95
Fully-Trainable CLIP	34.97	55.05	23.83

Table 4: Results of ablation experiments. The original HyCoLe-HNM is constructed with CLIP-ViT-B/32.

To probe the performance contribution from each design point of HyCoLe-HNM, we conduct the following five ablation experiments. As shown in Table 4, in each ablation experiment, we change the corresponding design point, and report the resulting overall performance on each dataset, which is the average value of the required R@Ks on the test

samples of that dataset.

- For the hybrid compositional learning mechanism, which includes both image compositional learning and text compositional learning, we disable text compositional learning by setting the loss scaling factor α to 0, which is applied to the TCM loss \mathcal{L}_{TCM} . As a result, we observe a slight performance drop on all datasets.
- For the heuristic negative mining method, which is based on heuristic rules and thus more efficient than hard negative mining methods, we replace it with a hard negative mining method. As a result, we observe a significant performance drop on FashionIQ and Fashion200K, and a slight one on MIT-States.
- For the gated fusion mechanism, which is borrowed from a QA model to implement the fusion module, we replace it with a simple addition operation. As a result, we observe a significant performance drop on all datasets.
- For privileged information, which is in the form of image-related texts and applied to cross-modal representation learning and text compositional learning, we disable its application by setting the loss scaling factors α and β to 0, which are separately applied to the TCM loss \mathcal{L}_{TCM} and the sum of the RITM loss \mathcal{L}_{RITM} and the TITM loss \mathcal{L}_{TITM} . As a result, we observe a slight performance drop on all datasets.
- For the fine-tuning of CLIP, which is controlled by the backbone activity ratio γ , we examine two extreme cases. On the one hand, we freeze CLIP by setting γ to 0. On the other hand, we make CLIP fully-trainable by setting γ to 1. As a result, we observe a significant performance drop on all datasets in both cases.

5 Conclusions

In this paper, we propose the CIR model HyCoLe-HNM, where we use the pre-trained V&L model CLIP as the backbone, utilize privileged information in the form of image-related texts to perform cross-modal representation learning and hybrid compositional learning, borrow a gated-fusion mechanism from a QA model to perform compositional fusion, and filter negative samples through



Figure 4: Retrieval examples from FashionIQ, Fashion200K, and MIT-States. For each example, the query image and the modification text are shown on the left, and the retrieved images are ranked by their matching scores and shown on the right with the target image highlighted.

heuristic negative mining. Experimental results show that HyCoLe-HNM achieves SOTA performance on three CIR datasets, namely FashionIQ, Fashion200K, and MIT-States. In the future, we plan to re-rank the top few candidate images retrieved by HyCoLe-HNM through certain cross-modal attention mechanisms, which we believe can further improve performance.

6 Limitations

Besides conducting experiments on FashionIQ, Fashion200K, and MIT-States, which are all comprised of natural images, we also conduct experiments on another CIR dataset CSS (Vo et al., 2019), which is comprised of synthetic images. However, on CSS, the performance of HyCoLe-HNM is inferior to that of TIRG (R@1: 67.3% vs 73.7%). Since

the backbone of HyCoLe-HNM is CLIP, which is pre-trained on natural image-text pairs, we conjecture that the reason behind this under-performance is the domain shift between the natural images used to pre-train CLIP and the synthetic images in CSS used to train HyCoLe-HNM.

Acknowledgments

We are grateful to all our colleagues in the Toronto AI Lab of LG Electronics for their support. Among them, we would especially like to thank **Manasa Bharadwaj** and **Kevin Ferreira** for their valuable inputs, and thank **Yolanda Liu** and **Parya Nejat** for their help with computing resources. Besides, we would also like to thank the anonymous reviewers and area chairs for their constructive efforts to help us improve the paper.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Muhammad Umer Anwaar, Egor Labintsev, and Martin Kleinstueber. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1140–1149.
- Yanbei Chen and Loris Bazzani. 2020. Learning joint visual semantic matching embeddings for language-guided retrieval. In *European Conference on Computer Vision*, pages 136–152. Springer.
- Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020a. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. 2020. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- Shiv Ram Dubey. 2021. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Chunbin Gu, Jiajun Bu, Zhen Zhang, Zhi Yu, Dongfang Ma, and Wei Wang. 2021. Image search with text feedback by deep hierarchical attention mutual information maximization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4600–4609.
- Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471.
- Judy Hoffman, Saurabh Gupta, and Trevor Darrell. 2016. Learning with side information through modality hallucination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 826–834.
- Mehrdad Hosseinzadeh and Yang Wang. 2020. Composed query image retrieval using locally features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3596–3605.
- Phillip Isola, Joseph J Lim, and Edward H Adelson. 2015. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391.
- Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021a. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, pages 1–9.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021b. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. 2018. Spigan: Privileged adversarial learning from simulation. *arXiv preprint arXiv:1810.03756*.
- Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812.
- Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. 2020. Learning with privileged information for efficient image super-resolution. In *European Conference on Computer Vision*, pages 465–482. Springer.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.
- Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantic aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.
- Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.
- Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 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*, pages 2125–2134.
- Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. 2021. Transformer reasoning network for image-text matching and retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5222–5229. IEEE.
- Ishan Misra, Abhinav Gupta, and Martial Hebert. 2017. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801.
- Taylor Mordan, Nicolas Thome, Gilles Henaff, and Matthieu Cord. 2018. Revisiting multi-task learning with rock: a deep residual auxiliary block for visual detection. *Advances in neural information processing systems*, 31.
- Tushar Nagarajan and Kristen Grauman. 2018. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185.
- Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 30–38.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*.
- Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2018. Deep shape matching. In *Proceedings of the european conference on computer vision (eccv)*, pages 751–767.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.
- Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. 2021. Stylemepup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8504–8513.
- Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12.
- Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.
- Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*.
- Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 982–997.

- Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, and Vicente Ordonez. 2019. Drill-down: Interactive retrieval of complex scenes using natural language queries. *Advances in neural information processing systems*, 32.
- Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.
- Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval—an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448.
- Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013.
- Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714.
- Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive linguistic-visual composition network for image retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1369–1378.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*.
- Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. 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*, pages 11307–11317.
- Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474.
- Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1520–1528.
- Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.
- Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. 2021. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2074–2084.

A Appendix

Statistics	FashionIQ	Fashion200K	MIT-States
Number of Training Samples	17965	42707540	9790710
Number of Training Images	25097	108366	43207
Number of Test Samples	6007	30960	2397080
Number of Test Images	8570	3356	10546
Average Length of Reference Text (words)	11.38	4.32	2.0
Average Length of Target Text (words)	11.38	4.32	2.0
Average Length of Modification Text (words)	10.66	5.0	5.0

Table 5: Datasets statistics of FashionIQ, Fashion200k and MIT-States.







Field	FashionIQ	Fashion200K	MIT-States
Reference Image			
Reference Text	is wash, long sleeve, clean, print shift, bell, scoop, tunic.	is green, seamed, a-line, dress.	ripe fig.
Target Image			
Target Text	is clean, wash, sheath, v-neck, sleeveless, stretch, zipper, york.	is pink, seamred, a-line, dress.	unripe fig.
Modification Text	is solid black with no sleeves, is black with straps.	is not green, is pink.	is not ripe, is unripe.

Table 6: Data examples of FashionIQ, Fashion200k and MIT-States.