Decomposing Semantic Shifts for Composed Image Retrieval

Xingyu Yang[1,2*], **Daqing Liu**[3], **Heng Zhang**[4], **Yong Luo**[1,2†], **Chaoyue Wang**[3], **Jing Zhang**[5]

[1]School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China,
[2]Hubei Luojia Laboratory, Wuhan, China,
[3]JD Explore Academy, JD.com, China,
[4]Gaoling School of Artifical Intelligence, Renmin University of China, China,
[5]School of Computer Science, The University of Sydney, Australia,
{yangxingyu2021,luoyong}@whu.edu.cn, liudq.ustc@gmail.com, zhangheng@ruc.edu.cn,
chaoyue.wang@outlook.com, jing.zhang1@sydney.edu.au

Abstract

Composed image retrieval is a type of image retrieval task where the user provides a reference image as a starting point and specifies a text on how to shift from the starting point to the desired target image. However, most existing methods focus on the composition learning of text and reference images and oversimplify the text as a description, neglecting the inherent structure and the user's shifting intention of the texts. As a result, these methods typically take shortcuts that disregard the visual cue of the reference images. To address this issue, we reconsider the text as instructions and propose a Semantic Shift Network (SSN) that explicitly decomposes the semantic shifts into two steps: from the reference image to the visual prototype and from the visual prototype to the target image. Specifically, SSN explicitly decomposes the instructions into two components: degradation and upgradation, where the degradation is used to picture the visual prototype from the reference image, while the upgradation is used to enrich the visual prototype into the final representations to retrieve the desired target image. The experimental results show that the proposed SSN demonstrates a significant improvement of 5.42% and 1.37% on the CIRR and FashionIQ datasets, respectively, and establishes a new state-of-the-art performance. Code is available at https://github.com/starxing-yuu/SSN.

1 Introduction

Composed Image Retrieval (Vo et al. 2019) (CIR) is an emerging image retrieval task in that the users can provide a multi-modal query composed of a reference image and a text. Different from the traditional image retrieval (Weinzaepfel et al. 2022) where the users must provide the exact same image of the desired result or text-to-image retrieval (Wang et al. 2019) where the users should describe the target in a detailed language, as shown in Figure 1a, CIR relax the requirement of input thus the users can simply provide an example image that similar to the desired image as reference, and then describe the difference from the reference to the target. Despite their diverse model architectures (Kim et al. 2021; Yang et al. 2021), the essence of this task is to fully understand the user's intent conveyed by the reference image and the language and then find the most similar image from all candidates. Thanks to the development of vision features (Radford et al. 2021) and language representations (Devlin et al. 2019), we can accomplish the CIR task with more flexible free-form languages, including changing one specific attribute of one object and adding or removing some objects.

However, capturing the user's intentions is still a challenging problem because the instruction text is far different

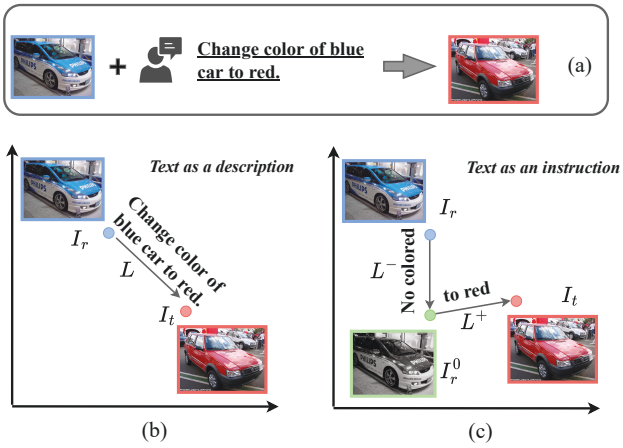

Figure 1: (a) gives an example of the CIR task. (b) shows existing works treat text as a description connecting the reference image and target image. They follow the paradigm of $I_r + L \leftrightarrow I_t$. (c) gives a brief illustration of our idea. We propose to consider the text as an instruction, inheriting the property of human language to express semantic shifts. With text instructions, the reference image is first degraded into visual prototypes and then enriched into the final representations to retrieve. This process can be described as $I_r \xrightarrow{L^-} I_r^0 \xrightarrow{L^+} I_t$.

*Contribution during internship at JD Explore Academy.
†Corresponding author.

from the description text that is commonly used in current vision-language tasks, e.g., visual grounding (Deng et al. 2018), cross-modal retrieval (Wang et al. 2019), or visual captioning (Liu et al. 2022a). For example, given the reference image and the text "Change the color of blue car to red" in Figure 1b, how to depict the desired image for us humans? One may have the following intuitive procedure: 1) identify the part of the reference image that should change, i.e., "the color of blue car", 2) imagine the visual prototype of the reference image, i.e., this car without any color, and 3) picture the final desired target image as "the car in red".

Unfortunately, despite the complex model architecture design with cross-modal attention (Hosseinzadeh and Wang 2020), graph neural network (Zhang et al. 2022), or fine-grained visual network (Hosseinzadeh and Wang 2020), existing methods (Baldrati et al. 2022; Zhao, Song, and Jin 2022; Goenka et al. 2022) generally oversimplify the CIR task as a composition learning of vision and language where the text is usually treated as a description (Figure 1b), disregarding the propriety of the structure of the text, which should be an instruction on how to modify the reference to the target. More seriously, composition learning typically introduces redundant or even incorrect information that may disrupt the final representations, e.g., simply combining the semantics 'blue, car' of the reference image and the semantics 'red, car, blue' of the text will depict 'a red and blue car' defectively. The fundamental cause of this issue lies in a lack of precise understanding of the language.

In this paper, we propose to take the text as instructions that represent the semantic shifts from the reference image to the target image, and then decompose the instructions into two parts: the degradation and the upgradation. As illustrated in Figure 1c, based on the decomposition, we conduct the desired image representations in two steps: 1) degrading the reference image into the visual prototype that only contains the visual attributes that need to be preserved, and 2) upgrading the visual prototype into the final desired target. Thanks to the decomposition of instructions, we divide the complex task that models the user's intentions into two simple and orthogonal sub-tasks which are easier to learn. Based on the final representations, we can directly find the nearest neighbors in the latent space as the final retrieval results.

Specifically, we implement the proposed method with a Semantic Shift Network (SSN) that is composed of four components: 1) the representation networks that extract visual and language features of reference images, target images, and instructions; 2) the decomposing network that decomposes the instruction text into the degradation part and upgradation part; 3) the degrading network that transforms the reference image to the visual prototype conditioned on the degradation part of the instruction text; 4) the upgrading network that transforms the visual prototype to the final representation of desired image conditioned on the upgradation part. To train the SSN, we design a traditional retrieval loss to guarantee the overall performance of composed image retrieval, as well as a regularization constraint that disciplines the language decomposing and the visual prototypes.

We validate the effectiveness of SSN on two widely used composed image retrieval benchmarks, i.e., FashionIQ (Wu et al. 2021) and CIRR (Liu et al. 2021). SSN stands as a new state-of-the-art on all metrics. Specifically, we achieve impressive improvements of 5.42% and 1.37% on CIRR and FashionIQ mean recall metrics, respectively.

In summary, our contributions include:

- We reformulate the composed image retrieval task as a semantic shift problem based on the text instructions, with the shift path as reference image $\rightarrow$ visual prototype $\rightarrow$ desired target image.
- We introduce a Semantic Shift Network for the CIR task that implements the decomposed semantic shifts with several well-designed components.
- The proposed SSN achieves state-of-the-art performance with impressive improvements on two widely-used composed image retrieval datasets.

2 Related Work

Image Retrieval. Image Retrieval is a fundamental task for the computer vision community since it has a wide range of application scenarios, e.g., search engines, and e-commerce (Zhang and Tao 2020). Given a query image, we need to return the most similar image. In the beginning, global image representation (Chen et al. 2022a) based retrieval methods were investigated. To achieve fine-grained matching (Sun et al. 2021) between images and thus improve retrieval performance, several approaches transformed images into several local representations (e.g., region features (Teichmann et al. 2019)). However, these pioneering works (Chen et al. 2022a; Sun et al. 2021; Weinzaepfel et al. 2022)'s queries are images only and they focus more on similarity matching between images. In reality, people usually convey query intent with text rather than images, so text-image retrieval is also a research focus in image search. Benefiting from the success of a large model for visual language pre-training (Kim, Son, and Kim 2021; Devlin et al. 2019), cross-modal representations have shown remarkable performance in text-image retrieval. Moreover, the query can also combine image and text, which is the direction we explore.

Composed Image Retrieval. Composed Image Retrieval refers to searching target images given semantically related reference images and modification texts. One line of work (Zhang et al. 2023; Kim et al. 2021) introduces target images into the forward process during training, which greatly increases the training cost. The MCL&SAP (Zhang et al. 2023) method perceives the semantics of modification texts at multiple layers of the image and models the differences in the image. Another line of works aiming at efficient retrieval explores composition learning and the following introduced works belong to this type. Gated residual fusion is first proposed to combine image and text features for the CIR task in TIRG (Vo et al. 2019) and is commonly used for global fusion in later works (Wang et al. 2022; Kim et al. 2021; Chen and Bazzani 2020). MAAF (Dodds et al. 2020) method applies self-attention mechanisms to realize the interaction between image-text sequences. Conditioned on text, VAL (Chen, Gong, and Bazzani 2020) is proposed to obtain combined features through multi-grained cross-modal semantic alignment, and CosMo (Lee, Kim, and Han

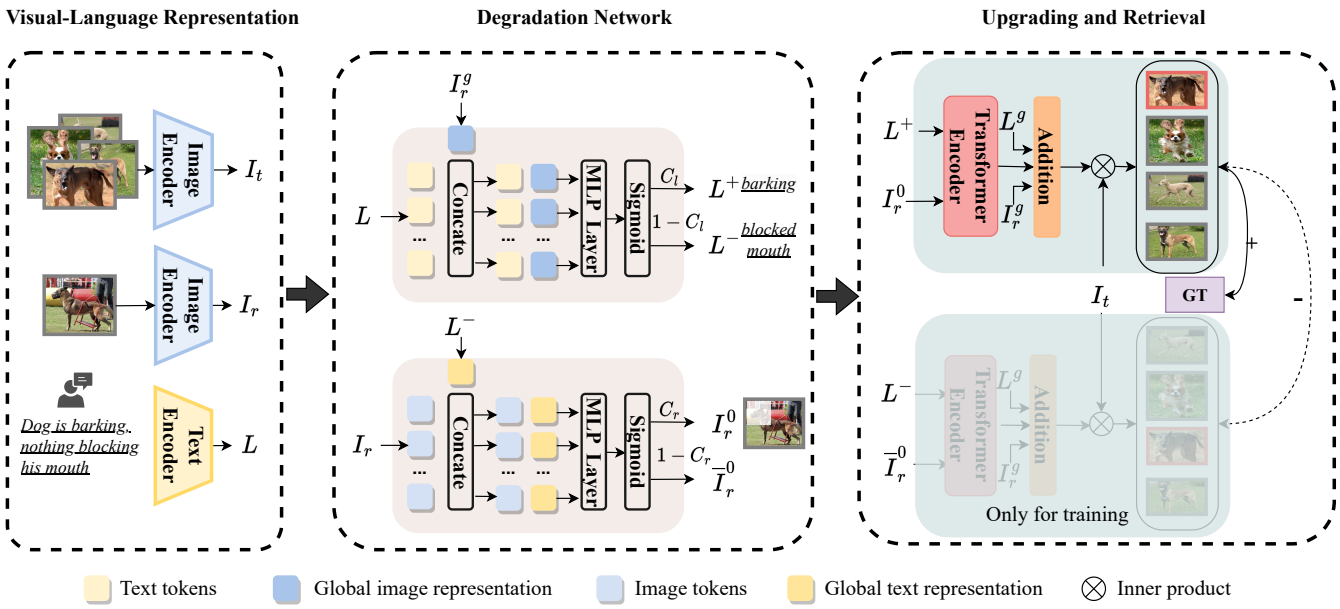

Figure 2: The pipeline of our proposed Semantic Shift Network (SSN). Given a pair of reference images and text modifiers (also as an instruction), we aim at retrieving the correct target image from candidate images. At the stage of visual-language representation, we utilize the CLIP image and text encoders to obtain the respective features. Then the semantic shift features from text instruction are decomposed to direct the reference image features I_r into a visual prototype I_r^0. The text features biased toward the target and reference image (namely L^+ and L^- respectively) are generated at the same time. In the upgrading process, L^+ is fused with visual prototypes I_r^0 by transformer an encode layer and then are linearly added to global representations. Finally, similarity scores are measured by an inner-product operation to generate the ranked list.

2021) method refines reference image features from the term of style and content. Benefiting from the visual-language pre-training, several works (Liu et al. 2021; Goenka et al. 2022; Saito et al. 2023; Baldrati et al. 2022) transfer to the CIR task and achieved favorable performance. Representative works include CLIP-based models (Baldrati et al. 2022; Saito et al. 2023) but the work in (Saito et al. 2023) is under the zero-shot setting, which is different from our task. Following (Baldrati et al. 2022), we also use the same CLIP encoder. Different from previous works, our model considers the modification text as an instruction that guides the reference image semantically shift back to a visual prototype.

3 Approach

In this section, we present the proposed Semantic Shift Network (SSN). We first briefly describe how to obtain the representations for image and text inputs. Then we present the technical details of degradation and upgradation, respectively, and finally depict our training objectives.

3.1 Preliminaries

The Composed Image Retrieval (CIR) task replaces the query in traditional image retrieval with multi-modal input, usually an image plus a text modifier. In this task, the query image is referred to as a reference image r, and the text modifier is denoted as l. Given each query $q = (r, l)$, the trained model returns a ranked list of the candidate images from a large image gallery $\mathcal{D}$, in descending order of similarity to the joint query semantic representation. An ideal retrieval system should rank the target image t at the first position.

3.2 Visual-Language Representation

CLIP (Radford et al. 2021) is a recently successful visual-language pre-trained model learned contrastively from 400M associated image-text pairs crawled from the internet. To leverage the powerful representation capability of CLIP, we adopt CLIP matched encoder to yield image and text features. Formally, we denote the CLIP image encoder as Φ_I and the CLIP text encoder as Φ_L. Given a triplet (r, l, t) of reference image r, modification text l, and target image t, image features $I_r = \Phi_I(r), I_t = \Phi_I(t)$ and text feature L is represented as $\Phi_L(l)$. Unlike previous works (Baldrati et al. 2022), we also preserve the fine-grained token-level features in addition to the global representations, which facilitates the exploration of richer interactions between modalities. We use a linear layer projecting image token-level features to the same d-dimension as text modality representations $(d = 512)$. Note that the features are a set of token-level features and projected global representations, they can be formulated in a unified way as follows:

$$V = \{\text{proj}(\boldsymbol{v}_{cls}), \boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_M\}, \quad (1)$$

where M is the sequence length or the number of tokens and tokens in V can be from r, l, t and thus produces I_r, I_t or L.

3.3 Degradation Process

We propose to model the CIR as a degradation-upgradation learning process where we treat the text as an instruction. The degradation process is to decompose features, during which the reference image is degraded into visual prototypes with decomposed text features guidance.

Inspired by the token selection (Liu et al. 2022b) in vision transformers, we propose a trainable cross-modal decompose module to direct the semantics towards the degradation part and upgradation part. With the help of reference images, we can distinguish this set of opposite semantic information. As shown in Figure 2, the inputs are a set of tokens from the provided text $\{\boldsymbol{v}_1^l, \boldsymbol{v}_2^l, ..., \boldsymbol{v}_M^l\} \in \mathbb{R}^{M\times d}$ (M is the text sequence length), we first concatenate them with the global semantic representation of the reference image $I_r^g = \text{proj}(\boldsymbol{v}_{cls}^r)$. Then we feed the concatenated features $X^l = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_M\}$, $\boldsymbol{x}_i = [I_r^g, \boldsymbol{v}_i^l] \in \mathbb{R}^{M\times 2d}$ ($[\cdot]$ is the concatenation operation) to one MLP-sigmoid layer. This reflects the weight of each token contribution to the degradation and upgradation semantic parts:

$$C_l = \text{sigmoid}(W_l X^l + b_l) \in \mathbb{R}^{M\times 1}. \tag{2}$$

Multiplied with the original token-level text features, we generate the positive and negative guiding text representations by Eq.(3). L^+ denotes the semantics related to the target, such as the expected attribute, "*red color*" (in Figure 1). Complementarily, L^- implies the object needed to be modified in the reference image, e.g., "*the car*" (in Figure 1). The conflicting property (red and blue color) is removed.

$$\begin{aligned} L^+ &= C_l \odot L \\ L^- &= (1 - C_l) \odot L, \end{aligned} \tag{3}$$

So far, we have obtained the guidance text features. Conditioned on this, to determine which among the reference images should be kept or discarded, we proceed with similar decomposition to produce contribution weights of reference image features but exchange the roles of the two modalities:

$$\begin{aligned} &X^r = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_P\}, \boldsymbol{x}_i = [L^-, \boldsymbol{v}_i^r] \in \mathbb{R}^{P\times 2d} \\ &C_r = \text{sigmoid}(W_r X^r + b_r) \in \mathbb{R}^{P\times 1}, \end{aligned} \tag{4}$$

where P is the number of patches and X^r is the concatenated features of reference image tokens $\{\boldsymbol{v}_1^r, \boldsymbol{v}_2^r, ..., \boldsymbol{v}_P^r\}$ and global semantic representations of L^-. After that, we obtain visual prototypes I_r^0 by

$$\begin{aligned} I_r^0 &= C_r \odot I_r \\ \overline{I}_r^0 &= (1 - C_r) \odot I_r, \end{aligned} \tag{5}$$

where I_r^0 preserves the core information of the given reference image. $\overline{I}_r^0$ is the token features after removing the visual prototypes from the original reference image.

3.4 Upgrading Process

Based on visual prototypes, the upgrading process aims to transform the visual prototype into the final representation close to the target image by compositional learning. In our late composition module, there are two parallel branches, one for processing positive guiding text and visual prototypes, and the other for negative guidance and irrelevant features in reference images. In each branch, we first add modality-specific embeddings $E_l \& E_i$ for inputs from different modalities, like modal-type embeddings in ViLT (Kim, Son, and Kim 2021). Note that these two learnable embeddings do not include degradation and upgrading information and they are used to indicate modality information. Then we fuse them via the transformer encode block $\mathcal{F}(\cdot)$. In detail, take the top branch as an example, given token sequences $[L^+, I_r^0]$, the fused features represent as:

$$F_{en} = \mathcal{F}([L^+ + E_l, I_r^0 + E_i]), \tag{6}$$

where E_l, E_i is the text modality embedding and image modality embedding respectively. The same fusion as in Eq.(6) is performed in the bottom branch for the inputs $\{L^-, \overline{I}_r^0\}$. Since the fusion layer takes sequence features as input, it can accept independent image token features rather than token features concatenated with text modality. Therefore we make the features of the target image also go through the fusion layer $\mathcal{F}$ to generate F_{tg} but without text inputs.

Finally following the work in (Baldrati et al. 2022), the final predicted feature F_p is a linear addition between the convex combination of the global reference image I_r^g and global text features L and the learned pooled fused features $\hat{F_{en}}$. We denote the final fused features in the top branch as F_p^+ and the fused features in the bottom branch as F_p^-.

3.5 Training and Inference

As shown in Figure 2, we use the inner product to measure the similarity between the predicted features and the target image representation and then obtain the ranking list of candidate images. Following (Baldrati et al. 2022; Zhao, Song, and Jin 2022; Wang et al. 2022), the retrieval objective is to minimize batched-based classification loss as follows:

$$\mathcal{L}_c = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp(\lambda * s(F_p^{(i)}, F_{tg}^{(i)}))}{\sum_{j=1}^{B}\exp(\lambda * s(F_p^{(i)}, F_{tg}^{(j)}))}, \tag{7}$$

where λ is a temperature parameter. Given the two sets of predicted features output by the late composite module, we obtain their similarity distribution to the target image. $\mathrm{z}^- = \text{softmax}(sim(F_p^-, F_{tg})), \mathrm{z}^+ = \text{softmax}(sim(F_p^+, F_{tg}))$. Finally, we employ a Kullback-Leibler Divergence loss as a regularization constraint:

$$\mathcal{L}_k = \text{KL}(\mathrm{z}^+ \| \mathrm{z}^{gt}) - \text{KL}(\mathrm{z}^- \| \mathrm{z}^+). \tag{8}$$

We aim to push away the distance between z^+ and z^- and thus optimize decompose learning. Thanks to the end-to-end training, the overall objective $\mathcal{L}$ guides the learning of L^+, L^- and I_r^0 and is described as follows:

$$\mathcal{L} = \mathcal{L}_c + w_{\mathcal{L}_k}\mathcal{L}_k, \tag{9}$$

where $w_{\mathcal{L}_k}$ is the hyperparameter of the loss weights, and its default value is 1. It is worth noting that the bottom branch in Figure 2 used for fusing L^- and $\overline{I}_r^0$ is only for training. During inference, we only use F_p^+, the composite features to retrieve the target image.

Method	Recall@K				$\text{Recall}_{\text{subset}}$@K			Average
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
TIRG (Vo et al. 2019)	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
TIRG+LastConv (Vo et al. 2019)	11.04	35.68	51.27	83.29	23.82	45.65	64.55	29.75
MAAF (Dodds et al. 2020)	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
MAAF+BERT (Dodds et al. 2020)	10.12	33.10	48.01	80.57	22.04	42.41	62.14	27.57
MAAF-IT (Dodds et al. 2020)	9.90	32.86	48.83	80.27	21.17	42.04	60.91	27.02
MAAF-RP (Dodds et al. 2020)	10.22	33.32	48.68	81.84	21.41	42.17	61.60	27.37
ARTEMIS (Delmas et al. 2022)	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
CIRPLANT (Liu et al. 2021)	15.18	43.36	60.48	87.64	33.81	56.99	75.40	38.59
CIRPLANT w/OSCAR (Liu et al. 2021)	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
CLIP4Cir (Baldrati et al. 2022)	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
SSN	**43.91**	**77.25**	**86.48**	**97.45**	**71.76**	**88.63**	**95.54**	**74.51**

Table 1: Comparisons with the state-of-the-art methods for composed image retrieval on the CIRR dataset. Here we show all Recall@K, $\text{Recall}_{\text{subset}}$@K and the average metrics. The average metric is the mean value of Recall@5 and $\text{Recall}_{\text{subset}}$@1. Our complete SSN model obtains significant improvement compared to other SOTA methods. The best results are in bold.

Method	Tops&Tees		Dress		Shirt		Average		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	mean
TIRG (Vo et al. 2019)	19.08	39.62	14.87	34.66	18.26	37.89	17.40	37.39	27.40
JVSM (Chen and Bazzani 2020)	13.00	26.90	10.70	25.90	12.00	27.10	11.90	26.60	19.25
VAL (Chen, Gong, and Bazzani 2020)	27.53	51.68	22.53	44.00	22.38	44.15	24.15	46.61	35.38
CoSMo (Lee, Kim, and Han 2021)	29.21	57.46	25.64	50.30	24.90	49.18	26.58	53.21	39.90
CLVC-Net (Wen et al. 2021)	33.50	64.00	29.85	56.47	28.75	54.76	30.70	58.41	44.56
SAC (Jandial et al. 2022)	32.70	61.23	26.52	51.01	28.02	51.86	29.08	54.70	41.89
DCNet (Kim et al. 2021)	30.44	58.29	28.95	56.07	23.95	47.30	27.78	53.89	40.84
MAAF (Dodds et al. 2020)	27.90	53.60	23.80	48.60	21.30	44.20	24.30	48.80	36.55
CIRPLANT (Liu et al. 2021)	21.64	45.38	17.45	40.41	17.53	38.81	18.87	41.53	30.20
ARTEMIS (Delmas et al. 2022)	29.20	54.83	27.16	52.40	21.78	43.64	26.05	50.29	38.17
MUR (Chen et al. 2022b)	37.37	68.41	30.60	57.46	31.54	58.29	33.17	61.39	47.28
CLIP4Cir (Baldrati et al. 2022)	41.41	65.37	33.81	59.40	**39.99**	60.45	38.32	61.74	50.03
SSN	**44.26**	**69.05**	**34.36**	**60.78**	38.13	**61.83**	**38.92**	**63.89**	**51.40**

Table 2: Comparisons with the state-of-the-art methods for composed image retrieval on the FashionIQ dataset. Here we show all Recall@10 and Recall@50 across all categories. Our complete SSN model outperforms other state-of-the-art methods on most of the metrics. The best result is in bold.

4 Experiments

4.1 Datasets and Metrics

CIRR Dataset (Liu et al. 2021) is the released dataset of open-domain for the CIR task. Each triplet consists of real-life images with human-generated modification sentences. The real-life images come from the popular NLVR^2 dataset (Suhr et al. 2018), which contains real-world entities with reasonable complexity. In 36,554 triplets, 80% are for training, 10% are for validation, and 10% are for evaluation.

FashionIQ Dataset (Wu et al. 2021) is a realistic dataset for interactive image retrieval in the fashion domain. Each query is composed of one reference image and two natural language descriptions about the visual differences of the target image. Following (Baldrati et al. 2022; Kim et al. 2021), we use the original evaluation split, which includes 5,373, 3,817, and 6,346 images for three specific fashion categories: Tops&Tees, Dresses, Shirts.

Metrics. Following previous works (Baldrati et al. 2022; Delmas et al. 2022; Zhao, Song, and Jin 2022), we employ Recall within top-K as the retrieval performance, which indicates the ratio of the ground-truth target image in the top-K ranking list that is correctly retrieved.

4.2 Implementation Details

We utilize the CLIP (Radford et al. 2021) model to initialize the image encoder with ViT-B/32. The hidden dimension of the 1-layer 8-head transformer encoder is set to 512. The temperature λ of the main retrieval loss (in Eq.(7)) is equal to 100. Note that for FashionIQ, we fix the image encoder after one training epoch and fine-tune the text encoder only. We adopt AdamW optimizer with an initial learning rate of 5e-5 to train the whole model. We apply the step scheduler to decay the learning rate by 10 every 10 epochs. The batch

Method	Recall@K				Recall_subset@K			(R@5+R_sub@1)/2
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
Baseline	42.62	76.7	87.06	**97.54**	68.98	86.73	94.16	72.84
SSN($\mathrm{I_r}, \mathrm{L}$)	43.34	76.97	87.4	97.2	72.18	88.42	**95.24**	74.575
SSN($\mathrm{I_r^0}, \mathrm{L}$)	44.43	77.44	86.92	96.96	71.63	88.21	94.88	74.535
SSN($\mathrm{I_r}, \mathrm{L^+}$)	43.46	**77.64**	87.51	97.42	71.9	87.9	95.12	74.77
SSN	**45.13**	77.49	**87.75**	97.32	**73.04**	**88.64**	95.17	**75.265**

Table 3: Ablation Studies of our SSN model with different components and various settings for decomposition outputs. We report all Recall@K, $\mathrm{Recall_{subset}@K}$, and the mean recall on the validation set of the CIRR dataset.

	shared	$\mathcal{L}_k$	R@K		R_sub@1	mean
			K=1	K=5		
1	✓	✓	43.77	77.30	71.66	74.91
2	×	×	44.32	77.42	71.92	74.67
3	×	✓	**45.13**	**77.49**	**73.04**	**75.27**

Table 4: Ablation experiments on loss function terms $\mathcal{L}_k$ and an exploration of whether two decomposing layers (one for images and one for text) share parameters. We report all metrics on the validation set of CIRR dataset.

size is set to 128 and the network is trained for 50 epochs. All experiments can be implemented with PyTorch on a single NVIDIA RTX 3090 Ti GPU.

4.3 Comparison with State-of-the-Arts

Results on CIRR dataset are presented in Table 1 for the test set. Our model which learns to decompose visual prototypes and semantic shift outperforms the state-of-the-art in all metrics. Compared to CLIP4Cir (Baldrati et al. 2022), a strong competitor that has recently successfully applied the CLIP model to the CIR task, our model outperforms it by 5.42% mean recall $(\mathrm{R@5} + \mathrm{R_{sub}@1})/2$ and increases up to 5.38% in Top-1 recall metrics. As shown in Table 1, we also outperform other methods by a large margin.

Results on FashionIQ dataset are reported in Table 2 for the validation set. Although the model is not improved as much as on the CIRR dataset, our proposed method achieves state-of-the-art results for all categories in most cases. Compared to the strongest method, we improved the mean recall by 1.37%. The limited improvement is due to the domain gap between the fashion data and the open domain CLIP, the small size of the data, and the specialization for fine-tuning the CLIP image encoder.

4.4 Ablation Studies

Model architecture. In order to demonstrate the contributions of individual components in our design, we first conducted experiments about ablated models. Moreover, we also explored several various settings for the decomposition outputs that are used during upgradation. Table 3 presents the detailed results on the validation set of the CIRR dataset. The different ablated models are as follows:

- **Baseline**: it is the model without any designed module.
- **SSN**($\mathrm{I_r}, \mathrm{L}$): it is the SSN model without degradation. That means the inputs for the upgradation are original dense tokens extracted from the visual and textual encoder.
- **SSN**($\mathrm{I_r^0}, \mathrm{L}$): it is the complete SSN model where visual prototypes and original text features are decomposition outputs.
- **SSN**($\mathrm{I_r}, \mathrm{L^+}$): it is the complete SSN model where the original reference image and positive guiding text features are decomposition outputs.
- **SSN**: it is the full SSN model where positive guiding text features ($\mathrm{L^+}$) rich the degraded visual prototypes ($\mathrm{I_r^0}$) during upgradation and then produces the final representations.

There are three following observations in Table 3: 1) the SSN($\mathrm{I_r}, \mathrm{L}$) model slightly outperforms the baseline by 0.72% in Recall@1 because of fine-grained token features. Our proposed method (SSN) achieves the best performance and gains a more significant improvement over the baseline model (2.51% in Recall@1). This highlights the effectiveness of decomposing semantic shifts into two steps. 2) SSN($\mathrm{I_r}, \mathrm{L^+}$) is comparable to SSN ($\mathrm{I_r}, \mathrm{L}$) model. This is because when only decomposing text instructions without generating visual prototypes, semantic shifts lack a well-acted object. 3) Based on the SSN ($\mathrm{I_r}, \mathrm{L}$) model, two models (SSN($\mathrm{I_r^0}, \mathrm{L}$) & SSN) picturing visual prototypes from reference images achieved further improvements up to 1.79%. This supports the motivation discussed in Section 1 that the visual cue of the reference images should not be disregarded.

Loss function. Our total training objective in Eq.(9) involves two aspects: the main retrieval loss and additional regularized loss $\mathcal{L}_k$. To demonstrate the effectiveness of the regularized constraint decomposing features, we performed ablation experiments on $\mathcal{L}_k$ in Eq.(9). Comparing the second and third rows in Table 4, we observe that the model with a regularized loss $\mathcal{L}_k$ performs better than the one without $\mathcal{L}_k$, despite only a slight improvement. This shows additional loss $\mathcal{L}_k$ helps to learn optimal decomposed features from original CLIP representations.

Shared parameter of decomposing layer for image and text? In Figure 2, we employ the same structure: one MLP layer followed by a sigmoid activation function, to decompose the semantic shifts and obtain visual prototypes. To explore whether components with the same structure can share parameters, we conduct additional experiments. From

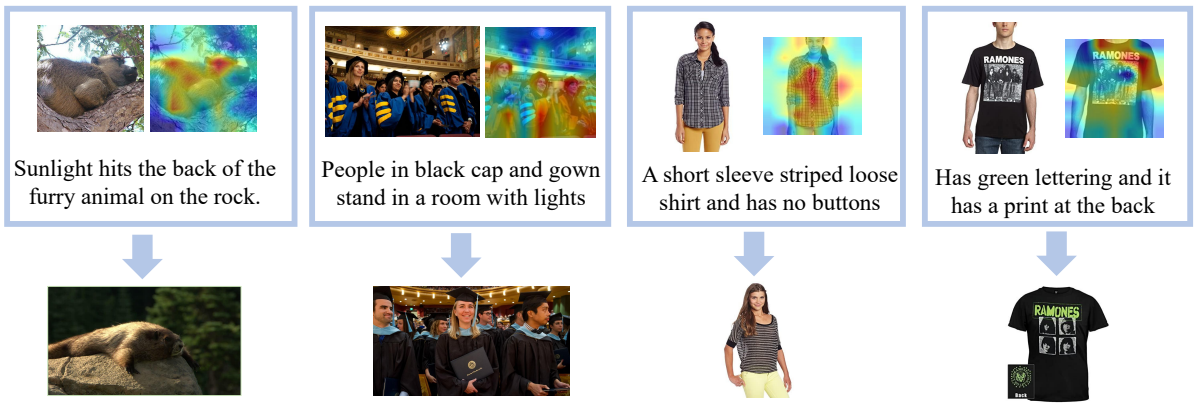

Figure 3: Visualization of where to be concerned when picturing visual prototypes from reference images on two datasets, in the form of heatmaps. In the reference image in the first column, the back of the furry animal is highlighted in the visual prototype, indicating the main characteristics of the target image. In the reference image of the fourth column, the reference image changes the print (lettering and color) in the front of the T-shirt to bring it back to the visual prototype. This suggests that we are not just concerned with the salient objects in the image and that the visual prototype contains a rich set of visual cues.

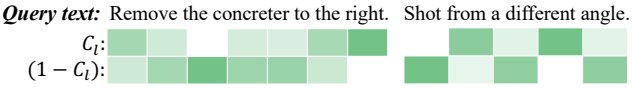

Figure 4: Examples of the learned C_l when decomposing text instructions to $L^+\&L^-$. The darker the green box, the higher the corresponding weight.

Figure 5: Top-4 retrieved results of the reference image, visual prototype and the proposed SSN.

rows one and three in Table 4, we can see that MLP layers that share parameters hurt the performance of the model. This is because decomposing networks for reference images and text have different objectives despite the same architecture. Producing visual prototypes is to get invariant features (shared with target images) while decomposing semantics in the text instruction is to get semantically shifted features.

4.5 Qualitative Results

Heatmaps in visual prototypes. As shown in Figure 3, we visualize what details are retained in the process of picturing visual prototypes from reference images on both datasets. In Section 3.3, we generate C_r to indicate whether certain features of the reference image are preserved or not. The heatmap is a merging of C_r and the original reference images. From the heatmaps, we can tell where the visual prototype and the original reference image have changed and the extent of these changes. With decomposing semantic shifts as guidance, we observe that the majority of important information in the reference image receives more attention.

Normalized weights in $L^+\&L^-$. We give examples of the learned weight C_l when decomposing text instructions to $L^+\&L^-$ in Figure 4. The word tokens with high weight in L^+ are those words representing semantic shifts, e.g., "*remove, to right*". While this type of words contribute little in L^-, those with high weights in L^- are some object words, corresponding to visual clues in the reference image.

Comparison of the retrieved results between different image inputs and SSN. From Figure 5, we see that the images retrieved by the reference image only are still "*Dog with dog*", while the images retrieved by the I_r^0 have removed another dog and preserved the most valuable cues, regardless of whether the dog was with something or not. Our SSN can put the correct image in the first position.

5 Conclusion

In this paper, we focus on the composed image retrieval task, an extended image retrieval task. Given the provided reference image and text requirements pair, the goal is to retrieve the desired target image. We first rethink the text as an instruction and then propose a Semantic Shift Network (SSN) to decompose the text instructions into degradation and upgradation. The text first directs the reference image toward the visual prototype and then guides the visual prototype closer to the target image. Extensive experiments on two benchmark datasets verify the effectiveness of the proposed method and show that our model significantly outperforms state-of-the-art methods by 5.42% and 1.37% on the mean of Recall@K, respectively. In the future, we intend to explore other complex mechanisms to model the text instruction in the CIR task and extend to the zero-shot setting.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (Grant No. U23A20318 and 62276195), Special Fund of Hubei Luojia Laboratory under Grant 220100014 and The Fundamental Research Funds for the Central Universities (No. 2042023kf1033).

References

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4959–4968.

Chen, W.; Liu, Y.; Wang, W.; Bakker, E. M.; Georgiou, T.; Fieguth, P.; Liu, L.; and Lew, M. S. 2022a. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, Y.; and Bazzani, L. 2020. Learning Joint Visual Semantic Matching Embeddings for Language-Guided Retrieval. In *European Conference on Computer Vision*.

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. 2022b. Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regularization. *arXiv preprint arXiv:2211.07394*.

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

Deng, C.; Wu, Q.; Wu, Q.; Hu, F.; Lyu, F.; and Tan, M. 2018. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7746–7755.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

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

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 *CVPR 2022*.

Hosseinzadeh, M.; and Wang, Y. 2020. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3596–3605.

Jandial, S.; Badjatiya, P.; Chawla, P.; Chopra, A.; Sarkar, M.; and Krishnamurthy, B. 2022. SAC: Semantic Attention Composition for Text-Conditioned Image Retrieval. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, 597–606. IEEE.

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, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

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.

Liu, B.; Wang, D.; Yang, X.; Zhou, Y.; Yao, R.; Shao, Z.; and Zhao, J. 2022a. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18041–18050.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022b. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, 319–335. Springer.

Liu, Z.; Opazo, C. R.; Teney, D.; and Gould, S. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV*, 2105–2114. IEEE.

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.

Saito, K.; Sohn, K.; Zhang, X.; Li, C.-L.; Lee, C.-Y.; Saenko, K.; and Pfister, T. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19305–19314.

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 preprint arXiv:1811.00491*.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8922–8931.

Teichmann, M.; Araujo, A.; Zhu, M.; and Sim, J. 2019. Detect-To-Retrieve: Efficient Regional Aggregation for Image Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Wang, C.; Nezhadarya, E.; Sadhu, T.; and Zhang, S. 2022. Exploring Compositional Image Retrieval with Hybrid Compositional Learning and Heuristic Negative Mining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 1273–1285.

Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; and Shao, J. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5764–5773.

Weinzaepfel, P.; Lucas, T.; Larlus, D.; and Kalantidis, Y. 2022. Learning super-features for image retrieval. *arXiv preprint arXiv:2201.13182*.

Wen, H.; Song, X.; Yang, X.; Zhan, Y.; and Nie, L. 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*, 1369–1378.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 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*, 11307–11317.

Yang, Y.; Wang, M.; Zhou, W.; and Li, H. 2021. Cross-modal Joint Prediction and Alignment for Composed Query Image Retrieval. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; César, P.; Metze, F.; and Prabhakaran, B., eds., *MM '21: ACM Multimedia Conference*, 3303–3311. ACM.

Zhang, F.; Yan, M.; Zhang, J.; and Xu, C. 2022. Comprehensive Relationship Reasoning for Composed Query Based Image Retrieval. In Magalhães, J.; Bimbo, A. D.; Satoh, S.; Sebe, N.; Alameda-Pineda, X.; Jin, Q.; Oria, V.; and Toni, L., eds., *MM '22: The 30th ACM International Conference on Multimedia*, 4655–4664. ACM.

Zhang, G.; Wei, S.; Pang, H.; Qiu, S.; and Zhao, Y. 2023. Enhance Composed Image Retrieval via Multi-level Collaborative Localization and Semantic Activeness Perception. *IEEE Transactions on Multimedia*.

Zhang, J.; and Tao, D. 2020. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*, 8(10): 7789–7817.

Zhao, Y.; Song, Y.; and Jin, Q. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1012–1021.