CoSMo: Content-Style Modulation for Image Retrieval with Text Feedback

Seungmin Lee* Dongwan Kim* Bohyung Han
Seoul National University
{dltmdals14, dongwan123, bhhan}@snu.ac.kr

Abstract

We tackle the task of image retrieval with text feedback, where a reference image and modifier text are combined to identify the desired target image. We focus on designing an image-text compositor, i.e., integrating multi-modal inputs to produce a representation similar to that of the target image. In our algorithm, Content-Style Modulation (CoSMo), we approach this challenge by introducing two modules based on deep neural networks: the content and style modulators. The content modulator performs local updates to the reference image feature after normalizing the style of the image, where a disentangled multi-modal non-local block is employed to achieve the desired content modifications. Then, the style modulator reintroduces global style information to the updated feature. We provide an in-depth view of our algorithm and its design choices, and show that it accomplishes outstanding performance on multiple image-text retrieval benchmarks. Our code can be found at: https://github.com/postBG/CosMo.pytorch

1. Introduction

Image retrieval is a crucial computer vision task that serves as the foundation for a variety of applications such as product search [22, 45, 61], person re-identification [69, 14, 42], and internet search [61, 52]. One of the most challenging aspects of building image retrieval systems is the ability to understand the user's intention accurately. Currently, a majority of image search engines are based on either image-to-image matching [52, 61] or image-text matching [59, 74, 70], where a user provides a single image or sentence as an input to find the most relevant images. However, it is not straightforward to express the complex target concept via a single image or text and design a model representing the intended concept. Furthermore, the users are unable to refine the retrieved results that fail to reflect their intention effectively.

We explore a different setting of image search—*image retrieval with text feedback* [66, 10]—where a *reference image* and a *modifier text* are used jointly as a query, as illustrated in Figure 1. Here, the reference image does not have any attribute labels, and the modifier text is a description of how the reference image should be changed to obtain the desired results, *i.e.*, *target images*. While inherently more complex than the standard image retrieval setting, this approach allows users to express their concepts more precisely by leveraging visual-linguistic information. In addition, since the users can recursively refine the search results based on previous results, the proposed algorithm would eventually lead to substantially improved output quality with high fidelity to input queries.

(a) Example of a **content** modification

(b) Example of a **style** modification

Figure 1. Examples of image retrieval with text feedback. Starting from a reference image, the user interacts with the system by providing a text input that expresses the desired changes. Given these inputs, the system retrieves images from a database that most accurately resembles the user's request. (a) and (b) depict examples of *content* and *style* modifications, respectively.

To tackle the task of image retrieval with text feedback, we create an image-text composition module that produces features similar to the target image features by combining the representations of the reference image and the modifier text. There are two main challenges in designing such a composition module. First, the module should be able to selectively preserve and modify the reference image features,

* equal contribution.

i.e. determine what to maintain and what to update. Second, the concepts conveyed in the modifier texts may range from being specific to certain *contents* of the image, to being more global and *stylistic* changes, as illustrated in Figure 1(a) and (b), respectively. As such, the module should be able to handle changes in both *content* and *style*.

A few works have contributed to the task of the image retrieval with text feedback. Most notably, TIRG [66] adopts gating and residual modules. The gating module uses a reference image and a modifier text to produce gate values that select what to update, while the residual module yields additive changes to the gated image feature. Although the algorithm design of TIRG [66] is intuitive, it fails to account for the wide range of input contents and styles, and thus, shows limited performance. On the other hand, VAL [10] employs multiple composition modules in varying depths of network, and each module produces the outputs that are used for retrieval. While VAL [10] is better suited to address the style-content issue, its use of additional modules in multiple layers demands much more resources.

In this work, we propose a novel image-text compositor, the **Co**ntent-**S**tyle **Mo**dulator (CoSMo), which directly addresses both the content and style changes conveyed by the modifier text. CoSMo consists of two modules: the *Content Modulator (CM)* and the *Style Modulator (SM)*. In CM, we introduce a Disentangled Multi-modal Non-local block (DMNL), which is an extension of the Non-local block [67] to the multi-modal setting, to effectively transform the contents of the reference image feature. To ensure that DMNL focuses on modifying contents rather than style, we first remove any style information by instance normalization of the reference image features. Moreover, we implement a few tricks in DMNL, which is imperative for stable training, especially in the multi-modal setting. In SM, we reintroduce style information to the transformed image feature. This module gates the channel-wise statistics of the original reference image feature, and predicts additional channel-wise statistics based on the modifier text feature.

Overall, our contributions are summarized as follows:

- We propose a novel image-text compositor, referred to as CoSMo, that is able to modulate both the contents and the style of the reference image.
- We design the content modulator, which employs the disentangled multi-modal non-local block with additional tricks to facilitate stable training. The proposed style modulator selectively preserves style information in the original image feature and adds new styles based on the text input.
- We demonstrate the effectiveness of CoSMo on multiple image-text retrieval benchmarks, where we outperform recent state-of-the-art methods. We also provide an extensive set of ablation experiments, as well as analysis that provide insight into our method.

2. Related Works

Image retrieval While traditional research on image search used handcrafted features [11], recent works have employed deep learning [52, 18, 61]. The most dominant image search engines are based on either image-text matching [59, 74, 70] or image-to-image matching [18, 52, 61], whereby users input a single image or sentence to find the most relevant images. Despite their simplicity, users are unable to refine any results that incorrectly reflect their intentions. Hence, many works have studied retrieval systems that take the user's feedback in various forms, such as spatial layouts [50, 48, 4], sketches [72, 13, 17, 57], attributes [75, 49, 2], or modifier texts [20, 10, 66]. In this paper, we focus on *image retrieval with text feedback*, since natural language is the most fundamental form of interaction between a user and system [20, 10].

To tackle this problem, we aim to develop a composite module that effectively integrates both image and text representation [10, 66, 27, 8]. To this end, TIRG [66] proposes a residual and gating module to compose image features with text features. While TIRG [66] is simple and intuitive, it fails to address the aforementioned style-content problem. To better address this issue, VAL [10] adopts multiple composition modules, and LBF [27] employs external off-the-shelf modules such as RPN [56]. While both methods are better suited for the task, using multiple composition modules or external modules requires extra resources. In contrast, our proposed method improves performance without using any off-the-shelf models or numerous composition modules by explicitly modulating the image's content and style to tackle the style-content problem.

Image as a combination of style and content Images have been interpreted as a combination of contents and style in various research areas, such as style transfer [76, 31, 16, 33, 38, 37], image synthesis [77, 41, 1, 35], and domain adaptation or generalization [7, 43, 26, 40, 60, 12]. Style is often referred to as the channel statistics that are spatially invariant, while contents are expressed by local features.

With this interpretation, previous works have shown that an image style can be modified while maintaining the content information [31, 43, 26, 76, 16, 38, 33, 37]. In many cases, the modification consists of normalization and modulation [31, 43, 60, 7]. For instance, AdaIN [31], a style transfer method, replaces the styles from an image feature using instance normalization (IN) [62]. Then, it modulates the normalized feature with the channel statistics of a target image. Likewise, in domain adaptation, AdaBN [43] first applies batch normalization (BN) [32] to an image from the source domain to remove the domain-specific style, then

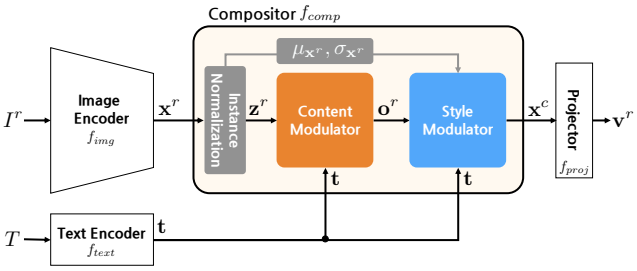


Figure 2. Overall pipeline of CoSMo.

modulates the image using the statistics of other domains.

Although the style-content interpretation is widespread, to the best of our knowledge, it has never been applied in our task. Following this interpretation, we normalize the reference image feature using IN and modulate its contents and styles using the proposed content modulator followed by style modulator.

Non-Local/Self-Attention mechanism The non-local or self-attention mechanism is an effective tool to capture long-range dependencies and context information between features in two different locations. Ever since the self-attention mechanism was first introduced for natural language processing tasks [64, 6], it has been adopted in various vision tasks such as classification [29], object detection [68, 19, 9], action recognition [67]. Moreover, many vision-language tasks such as image captioning [34, 3, 30] or VQA [73, 15, 46, 47, 51] have adopted the attention mechanism in various forms, *e.g.*, co-attention and intra-inter latent attention [39, 15, 51]. In the field of image retrieval with text feedback, LBF [27] applies the attention in a form similar to co-attention, but there is limited due to its use of an off-the-shelf RPN [56]. VAL [10] also employs the self-attention by concatenating the image and text features in each spatial location. However, we empirically verify that naïvely applying this form of self-attention to our method may lead to training instability. Inspired by the disentangled non-local block [71], we proposed a disentangled multi-modal non-local (DMNL) block that alleviates this instability. Hence, we utilize the DMNL as a core component of the content modulator for more stable learning and higher performance.

3. Proposed Method: Overview

Given an input query (I^r, T) based on a reference image, I^r, and a modifier text, T, we aim to generate image-text joint features that are well-aligned with the representation of the target image I^t. Our image-text composition framework consists of three major components: 1) the image encoder, 2) the text encoder , and 3) the image-text compositor.

The image and text encoders, denoted by $f_{\text{img}}(\cdot)$ and $f_{\text{text}}(\cdot)$, respectively, extract proper representations from the multi-modal inputs as

$$\mathbf{x}^r = f_{\text{img}}(I^r) \tag{1}$$
$$\mathbf{t} = f_{\text{text}}(T), \tag{2}$$

where $\mathbf{x}^r \in \mathbb{R}^{C\times(H\times W)}$ and $\mathbf{t} \in \mathbb{R}^N$ are the encoded reference image and the modifier text features, respectively. Then, the image-text compositor, $f_{\text{comp}}(\cdot,\cdot)$, transforms the image feature using the modifier text representations appropriately, which is given by

$$\mathbf{x}^c = f_{\text{comp}}(\mathbf{x}^r, \mathbf{t}). \tag{3}$$

The transformed feature $\mathbf{x}^c$ should be similar to the representation of the target image, which is given by $\mathbf{x}^t = f_{\text{img}}(I^t)$. To learn the model realizing the constraint, we project the image-text and target features onto an embedding space using a projection layer, which is given by

$$\mathbf{v}^c = f_{\text{proj}}(\mathbf{x}^c) \tag{4}$$
$$\mathbf{v}^t = f_{\text{proj}}(\mathbf{x}^t), \tag{5}$$

and compute the loss between the two projected features using the dissimilarity based on a distance metric. To mitigate any biases induced by the differences in vector norm, we include ℓ_2 normalization in the final projection layer.

Following TIRG, we use a batch-based classification loss (BBCL), which, compared to the triplet loss [25], achieves better discriminativeness and faster convergence in complex datasets [66]. Each batch consists of B pairs of query samples—a reference image and a modifier text—and their respective target images. The loss function is defined by

$$\mathcal{L}_{\text{BBCL}} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp\left(\kappa(\mathbf{v}^{c,i}, \mathbf{v}^{t,i})\right)}{\sum_{j=1}^{N}\exp\left(\kappa(\mathbf{v}^{c,i}, \mathbf{v}^{t,j})\right)}, \tag{6}$$

where $\kappa(\cdot,\cdot)$ is an arbitrary distance metric, *e.g.*, cosine distance. Finally, for image retrieval, we rank the distance to the projected features of the samples in the database from a composed image-text query feature.

Since the implementation of the encoder modules and the loss function is straightforward, we primarily focus on the compositor, *i.e.*, how to effectively fuse the text and image features. The next section describes the details of the proposed compositor with the intuition of our design choice.

4. Compositor Design

The proposed compositor contains two distinct modules—Content Modulator (CM) and Style Modulator (SM)—following the content-style interpretation [31, 43, 16, 33]. As their names imply, CM and SM modulate content and style of an input image based on the corresponding text input.

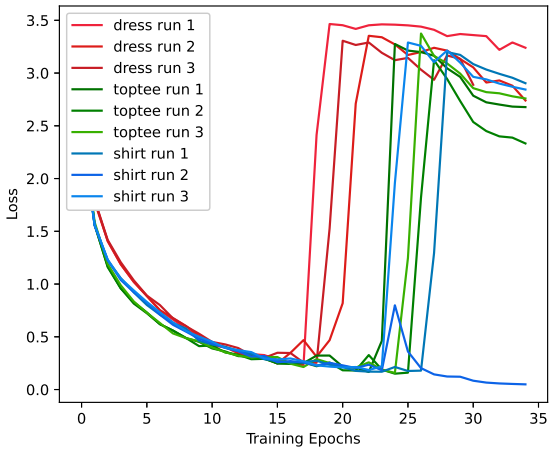


Figure 3. Unstable training of the naïve multi-modal non-local block. We run three independent runs for each sub-dataset in FashionIQ and observe that the training loss often diverges mid-training.

Figure 2 illustrates the overall pipeline to generate the image-text joint feature $\mathbf{x}^c$. In the compositor, we first calculate the underlying style information represented by $(\mu_{\mathbf{x}^r}, \sigma_{\mathbf{x}^r})$ and remove it from the image feature $\mathbf{x}^r$ by applying instance normalization, which is given by

$$\mathbf{z}^r = \mathrm{IN}(\mathbf{x}^r) = \frac{\mathbf{x}^r - \mu_{\mathbf{x}^r}}{\sigma_{\mathbf{x}^r}}. \tag{7}$$

Then, the contents and style are modified based on the text via CM followed by SM.

4.1. Content Modulator

The Content Modulator (CM) takes the image feature after instance normalization, $\mathbf{z}^r$, and text feature, $\mathbf{t}$, as its inputs and fuses them to generate a transformed feature $\mathbf{o}^r$, which is given by

$$\mathbf{o}^r = \mathrm{CM}(\mathbf{z}^r, \mathbf{t}) = \mathrm{conv}_{1\times1}(\mathbf{y}^r) + \mathbf{z}^r, \tag{8}$$

where $\mathrm{conv}_{1\times1}(\cdot)$ denotes a 1×1 convolution, and $\mathbf{y}^r \in \mathbb{R}^{C\times(H\times W)}$ is the output feature map of the Disentangled Multi-modal Non-Local block (DMNL), in which the non-local block [67] is appropriately redesigned to better suit our task. The residual connection allows the CM to modify the input feature, rather than completely replacing it with the output of DMNL. Additionally, by providing instance normalized image features, we allow the content modulator to focus solely on modifying the contents of the image, rather than the style.

4.1.1 Multi-modal Non-Local Block

We first outline the multi-modal non-local block (MNL) [67, 10]. We employ the MNL to capture the long-range dependencies between two positions of visual-linguistic features using the pairwise relationships. Unlike the original non-local block [67], MNL takes $\mathbf{z}^r$ and $\mathbf{t}$ and outputs a composed image-text feature, $\mathbf{y}_i^r \in \mathbb{R}^C$, which is given by

$$\mathbf{y}_i^r = \sum_{j\in\Omega} \omega(\mathbf{z}_i^r, \mathbf{z}_j^r, \mathbf{t}) g([\mathbf{z}_j^r, \mathbf{t}]), \tag{9}$$

where $\mathbf{z}_i^r$ and $\mathbf{z}_j^r$ are the C-dimensional vector obtained from $\mathbf{z}^r$ at the i^{th} and j^{th} locations, respectively, $\mathbf{t} \in \mathbb{R}^N$ is the encoded text feature, Ω indicates the set of all locations, and $[\cdot,\cdot]$ denotes the concatenation of two vectors.

As presented in Eq. (9), two functions make up the MNL block. The *value function*, denoted by $g(\cdot)$ is implemented by a simple multi-layer perceptron (MLP) while $\omega(\cdot,\cdot,\cdot)$ is the *similarity function*, which is given by

$$\omega(\mathbf{z}_i^r, \mathbf{z}_j^r, \mathbf{t}) = \psi(\mathbf{q}_i^T \mathbf{k}_j) = \frac{\exp(\mathbf{q}_i^T \mathbf{k}_j)}{\sum_{s\in\Omega} \exp(\mathbf{q}_i^T \mathbf{k}_s)}, \tag{10}$$

where $\psi(\cdot)$ is the softmax function. Note that $\mathbf{q}_i \in \mathbb{R}^M$ indicates a query and $\mathbf{k}_j \in \mathbb{R}^M$ is a key, which are further defined by

$$\begin{aligned} \mathbf{q}_i &= \mathbf{W}_q[\mathbf{z}_i^r, \mathbf{t}] = \mathbf{W}_{q_z}\mathbf{z}_i^r + \mathbf{W}_{q_t}\mathbf{t}, \\ \mathbf{k}_j &= \mathbf{W}_k[\mathbf{z}_j^r, \mathbf{t}] = \mathbf{W}_{k_z}\mathbf{z}_j^r + \mathbf{W}_{k_t}\mathbf{t}, \end{aligned} \tag{11}$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are learnable parameters.

Demystifying the similarity function To better understand the similarity function, we unravel the query-key dot product as

$$\begin{aligned} \mathbf{q}_i^T \mathbf{k}_j =& (\mathbf{W}_{q_z}\mathbf{z}_i^r + \mathbf{W}_{q_t}\mathbf{t})^T \cdot (\mathbf{W}_{k_z}\mathbf{z}_j^r + \mathbf{W}_{k_t}\mathbf{t}) \\ =& \underbrace{(\mathbf{W}_{q_z}\mathbf{z}_i^r)^T(\mathbf{W}_{k_z}\mathbf{z}_j^r)}_{A_j} + \underbrace{(\mathbf{W}_{q_t}\mathbf{t})^T(\mathbf{W}_{k_z}\mathbf{z}_j^r)}_{B_j} \\ &+ (\mathbf{W}_{q_z}\mathbf{z}_i^r)^T(\mathbf{W}_{k_t}\mathbf{t}) + (\mathbf{W}_{q_t}\mathbf{t})^T(\mathbf{W}_{k_t}\mathbf{t}), \end{aligned} \tag{12}$$

where $(\mathbf{W}_{q_z}\mathbf{z}_i^r)^T(\mathbf{W}_{k_z}\mathbf{z}_j^r) = A_j$ and $(\mathbf{W}_{q_t}\mathbf{t})^T(\mathbf{W}_{k_z}\mathbf{z}_j^r) = B_j$, for simplicity. When a softmax function is applied to Eq. (12), the last two terms are cancelled, which leads to

$$\begin{aligned} \omega(\mathbf{z}_i^r, \mathbf{z}_j^r, \mathbf{t}) &= \psi(\mathbf{q}_i^T \mathbf{k}_j) \\ &= \psi\left(A_j + B_j\right) \\ &= \lambda_i \psi\left(A_j\right) \cdot \psi\left(B_j\right) \\ &= \lambda_i \omega_s(\mathbf{z}_i^r, \mathbf{z}_j^r) \cdot \omega_c(\mathbf{t}, \mathbf{z}_j^r), \end{aligned} \tag{13}$$

where λ_i is a normalization constant. Refer to our supplementary document for details about the derivation. Here we observe that the similarity function, $\omega(\cdot,\cdot,\cdot)$, can be decomposed into the pixel-wise self-attention and the text-pixel cross-attention, denoted by $\omega_s(\cdot,\cdot)$ and $\omega_c(\cdot,\cdot)$, respectively.

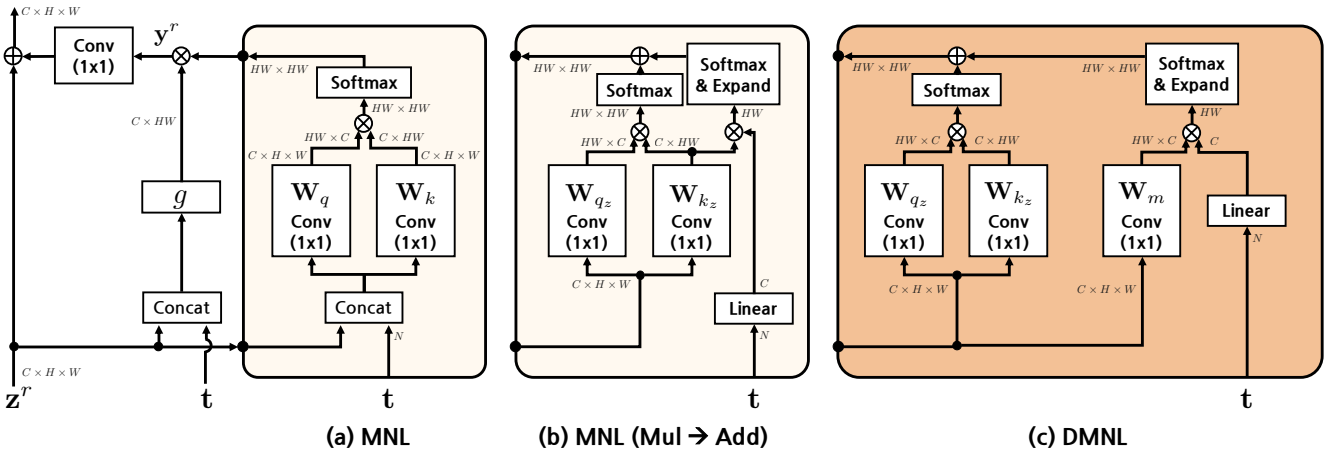


Figure 4. Difference between the three attention blocks according to the tricks introduced in 4.1.2. Starting from the multi-modal non-local (MNL) block (a), we replace the multiplication interaction with an additive one (b). Finally, we replace W_{k_z} with W_m in the text-pixel cross-attention (c).

Instability of the MNL block In practice, naïve application of the MNL block to our task results in unstable training as presented in Figure 3. This is partly because the gradients with respect to the pixel-wise self-attention and the pixel-text cross-attention are highly interdependent, which is derived from Eq. (13) as follows:

$$\frac{\partial \mathcal{L}}{\partial \omega_s} = \lambda_i \frac{\partial \mathcal{L}}{\partial \omega} \cdot \omega_c, \quad \frac{\partial \mathcal{L}}{\partial \omega_c} = \lambda_i \frac{\partial \mathcal{L}}{\partial \omega} \cdot \omega_s, \tag{14}$$

where $\mathcal{L}$ is an arbitrary loss function. Note that ω_s and ω_c often output values close to zero, which hinders gradient flows on its counterpart. This is particularly exacerbated in our task setting, where ω_s and ω_c are functions for different input modalities.

4.1.2 Disentangled Multi-modal Non-Local block

We alleviate this gradient entanglement issue by employing the following two tricks. First, we replace the multiplicative interaction between ω_s and ω_c with an additive one, as introduced in [71]. Specifically, we update Eq. (13) as

$$\hat{\omega}(\mathbf{z}_i^r, \mathbf{z}_j^r, \mathbf{t}) = \lambda_i \left(\omega_s(\mathbf{z}_i^r, \mathbf{z}_j^r) + \omega_c(\mathbf{t}, \mathbf{z}_j^r)\right). \tag{15}$$

By doing so, the gradients of ω_s and ω_c are no longer interdependent as shown below:

$$\frac{\partial \mathcal{L}}{\partial \omega_s} = \lambda_i \frac{\partial \mathcal{L}}{\partial \hat{\omega}}, \quad \frac{\partial \mathcal{L}}{\partial \omega_c} = \lambda_i \frac{\partial \mathcal{L}}{\partial \hat{\omega}}. \tag{16}$$

The second trick is related to further disentangling any shared parameters between the image and text features. In particular, the key projection parameter $\mathbf{W}_{k_z}$ is used in both the pixelwise self-attention and the text-pixel cross-attention terms (Figure 4(b)). With similar motivations to the first trick, we replace $\mathbf{W}_{k_z}$ in the text-pixel cross-attention with a new key projection matrix, $\mathbf{W}_m$:

$$\omega_c^*(\mathbf{t}, \mathbf{z}_j^r) = \psi\left((\mathbf{W}_{q_t}\mathbf{t})^T(\mathbf{W}_m \mathbf{z}_j^r)\right). \tag{17}$$

Then, the final similarity function for our DMNL block is given by

$$\omega_{\text{DMNL}}(\mathbf{z}_i^r, \mathbf{z}_j^r, \mathbf{t}) = \lambda_i \left(\omega_s(\mathbf{z}_i^r, \mathbf{z}_j^r) + \omega_c^*(\mathbf{t}, \mathbf{z}_j^r)\right). \tag{18}$$

Our DMNL block, illustrated in Figure 4(c), is an effective means of combining text and image features while maintaining stability in training. As with the attention block [64], the DMNL block can be implemented with numerous heads and can also be stacked multiple times to improve performance.

4.2. Style Modulator

The Style Modulator (SM) applies affine transformations to individual channels of $\mathbf{o}^r$, the output of our CM module. It calculates the affine parameters, $\gamma \in \mathbb{R}^C$ and $\beta \in \mathbb{R}^C$ by the following operations:

$$\begin{aligned} \gamma &= \text{sigmoid}\left(\phi_\gamma(\mathbf{t})\right) \cdot \sigma_{\mathbf{x}^r} + f_\gamma(\mathbf{t}), \\ \beta &= \text{sigmoid}\left(\phi_\beta(\mathbf{t})\right) \cdot \mu_{\mathbf{x}^r} + f_\beta(\mathbf{t}), \end{aligned} \tag{19}$$

where $\phi, f : \mathbb{R}^N \rightarrow \mathbb{R}^C$ are simple linear transformations. Note that $\mu_{\mathbf{x}^r}$ and $\sigma_{\mathbf{x}^r}$ are channel-wise statistics obtained from the reference image feature, $\mathbf{x}^r$, and encode its original style information [31, 43]. Two gating functions based on the text feature, ϕ_γ and ϕ_β, selectively preserve certain styles in the original feature while discarding others. Then, $f_\gamma(\cdot)$ and $f_\beta(\cdot)$ inject new style information based on the text feature.

Method	Dress		Toptee		Shirt		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Relationship [58]	15.44	38.08	21.10	44.77	18.33	38.63	18.29	40.49
MRN [36]	12.32	32.18	18.11	36.33	15.88	34.33	15.44	34.28
FiLM [55]	14.23	33.34	17.30	37.68	15.04	34.09	15.52	35.04
TIRG [66]	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39
VAL (single-level) [10]	-	-	-	-	-	-	20.53	42.57
VAL (multi-level) [10]	21.12	42.19	25.64	49.49	21.03	43.44	22.60	45.04
Ours	**25.64** $\pm$ 0.21	**50.30** $\pm$ 0.10	**29.21** $\pm$ 0.12	**57.46** $\pm$ 0.16	**24.90** $\pm$ 0.25	**49.18** $\pm$ 0.27	**26.58**	**52.31**

Table 1. Results on FashionIQ using ResNet-50.

The resulting affine parameters are employed to modulate the channel-wise statistics of $\mathbf{o}^r$ by the following equation, which corresponds to updating the style of the reference image:

$$\mathbf{x}^c = \gamma \mathbf{o}^r + \beta. \tag{20}$$

5. Experimental Results

We demonstrate the effectiveness of the proposed approach, CoSMo, by evaluating on several datasets. Tables 1, 2, and 3 present our main results, where we measure the Recall@K performance on the validation set of each dataset. All compared methods in each table adopt the same base encoders, unless stated otherwise. We run all experiments three times independently and report the averages with the maximum deviations from the averages. Furthermore, we provide qualitative results in the supplementary material.

5.1. FashionIQ

The FashionIQ [21] dataset is a commonly used benchmark for image retrieval with natural language feedback. It consists of 77,684 fashion product images that are split into three distinct categories: *Dress*, *Toptee*, and *Shirt*. The dataset is organized by triplets, with a reference image, a target image, and a pair of relative captions that describe the differences between the two images. For simplicity, we refer to the reference image and the relative caption as a query pair. Across the three sub-datasets, there are around 18,000 query pairs to train on. On FashionIQ, we adopt ResNet-50 [24] as the base image encoder.

One detail we are obligated to point out is that in our experiments, we use a slightly different evaluation split than the one defined by the original authors of the datastet [21]. We follow the evaluation method of VAL [9], where the evaluation split is defined by the union of reference and target images. Thus, compared to the original evaluation split, there are fewer test images, which leads to slightly higher performance than the original version. We evaluate with this reduced test set for fair comparison with other methods, including VAL. Additionally, we present results using the original evaluation split in our supplementary material.

Table 1 presents our results on the FashionIQ dataset, which are based on the reduced evaluation split. Compared to TIRG, on average, CoSMo achieves significantly improved performance—about 10% and 15% points for Recall@10 and Recall@50 metrics, respectively. Even when comparing to single-level VAL, we observe around 6% and 10% points better accuracy. Finally, we find it remarkable that CoSMo, despite being a single-level method, outperforms multi-level VAL by a significant margin, more than 3% points for Recall@10 and 7% points for Recall@50, on all three subcategories.

5.2. Shoes

The Shoes [5] dataset was originally proposed for attribute discovery, but has been additionally labeled for dialog-based interactive image retrieval [20]. It consists of 10,000 training queries and 4,658 validation examples. We employ ResNet-50 as the image encoder.

According to our observation, the results on the Shoes dataset have similar patterns with the ones on FashionIQ as presented in Table 2, albeit a slightly lower margin of increase. While the proposed approach is slighly behind multi-level VAL on the Recall@10 metric, it still outperforms multi-level VAL by 0.23% and 2.11% points on Recall@1 and Recall@50, respectively. We recognize the significance in these results since the results on VAL are outside our range of uncertainty in both metrics.

5.3. Fashion200K

The Fashion200K [23] is a large-scale dataset with over 200,000 fashion images crawled from multiple websites. While each image is labeled with various types of information including product information, review, and bounding boxes for clothes, we only utilize the raw images and their corresponding description. The description themselves are a list of attributes, such as *"multicolor french lace crew neck lace dress"* or *"black crepe tie front dress"*. Following [66], we convert these attributes into relative descriptions in an online-fashion (see supplementary material for additional information). Since retrieval is performed by matching the target attribute rather than the target image as in the Fash-

Method	Shoes		
	R@1	R@10	R@50
Relationship [58]	12.31	45.10	71.45
MRN [36]	11.74	41.70	67.01
FiLM [55]	10.19	38.89	68.30
TIRG [66]	12.60	45.45	69.39
VAL (single-level) [10]	14.20	46.65	-
VAL (multi-level) [10]	16.49	**49.12**	73.53
Ours	**16.72** ± 0.20	48.36 ± 0.12	**75.64** ± 0.41

Table 2. Results on Shoes using ResNet-50.

Method	Fashion200k		
	R@1	R@10	R@50
Param Hashing [53]	12.2	40.0	61.7
Show and Tell [65]	12.3	40.2	61.8
Relationship [58]	13.0	40.5	62.4
FiLM [55]	12.9	39.5	61.9
MRN [36]	13.4	40.0	61.9
TIRG [66]	14.1	42.5	63.8
LBF [27]	17.8	48.4	68.5
VAL (single-level) [10]	15.6	44.8	-
VAL (multi-level) [10]	21.2	49.0	68.8
Ours	**23.3** ± 0.3	**50.4** ± 0.2	**69.3** ± 0.2
JVSM* [8]	19.0	52.1	70.0

Table 3. Results on Fashion200k using ResNet-18. *JVSM uses additional labels during training.

ionIQ and Shoes datasets, multiple correct answers may exist. In total, the training split contains around 172,000 images and the test set contains 33,480 queries. For fair comparison with previous works, we use ResNet-18 as our image encoder on Fashion200K.

The results on Fashion200K further validate the effectiveness of CoSMo in this task. As mentioned earlier, Fashion200K is evaluated by matching the target features, and, consequently, the target image is not unique. The overall scores on Fashion200K are quite high, despite its much larger size and greater diversity compared to the FashionIQ and Shoes datasets.

Our results on Fashion200K are shown in Table 3. Here, we outperform TIRG by roughly 7%, 8%, and 6% points on the Recall@1, Recall@10, and Recall@50 metrics respectively. We also outperform LBF [27]—a method that employs an off-the-shelf region proposal network [56]—by a significant margin as well. We also show improvements over multi-level VAL, although we point out that VAL uses MobileNet-v1 [28][0] as their base image encoder. Finally, compared to JVSM [8], we observe stronger performance on Recall@1, but slightly weaker performance on Recall@10 and Recall@50. However, we note that JVSM employs additional label information during training.

5.4. Implementation Details

We evaluate CoSMo using two different image encoders, f_{img}: ResNet-18 and ResNet-50. Specifically, f_{img} is given by the output from layer 4 of the backbone networks. For ResNet-50, this layer will output a feature map with 2,048 channels, and for ResNet-18, there are 512 channels. The final projector, f_{proj}, consists of a Global Average Pooling layer followed by a linear layer, which projects onto a 512-dimensional vector. The output vector is ℓ_2 normalized and scaled by a factor of 4 for more efficient training.

Our text encoder, f_{text}, is composed of an embedding layer and an LSTM, followed by a single linear layer. Thus, we first embed the text into a 512-dimensional vector, transform it using an LSTM with 1,024 hidden neurons, and finally obtain $\mathbf{t} \in \mathbb{R}^{512}$ by the projection.

To train our model, we use a rectified Adam [44] optimizer with a base learning rate of 2×10^{-4}, which decays once after 30 epochs by a factor of 10. We train for a total of 80 epochs, with a batch size of 32 on ResNet-50 and 128 on ResNet-18. Our framework of choice is PyTorch [54].

[0]While the difference in model architectures may play a role in overall results, we note that MobileNet-v1 and ResNet-18 have almost identical scores on both ImageNet Top1 and Top5 error rates.

6. Analysis

We conduct in-depth analyses to help us better understand the inner workings of CoSMo and its two modules.

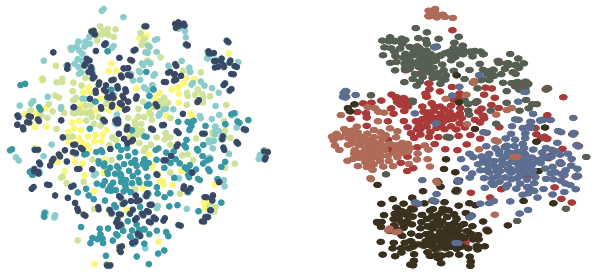
Role of CM and SM To measure the contributions of CM and SM, we conduct ablation experiments on each module. To evaluate the *CM only* version, we remove the IN layer to retain style information. By the design update, the DMNL blocks are responsible for modifying both the contents and style, which is challenging for a module designed to modify the contents only. To test the *SM only* case, we simply remove the content modulator.

The results of these ablations are presented in rows 4 and 5 of Table 4. As depicted in the table, *SM only* demonstrates stronger performance than TIRG and comparable performance to single-level VAL (see Table 1 for comparisons) despite lacking any non-local mechanism. Moreover, we observe that *CM only* shows even better performance than multi-level VAL on the FashionIQ dataset. Thus, the each module functions can be used as an effective standalone composition module. However, using CM and SM together demonstrates even stronger performance, which implies that explicitly tackling the style-content issue is reasonable for this task.

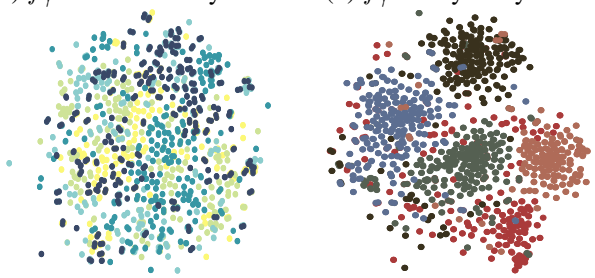
Effects of the two tricks in DMNL (4.1.2) Figure 3 illustrates the training instability resulting from a naive imple-

Method	FashionIQ		Shoes	
	R@10	R@50	R@1	R@10
MNL	N/A	N/A	N/A	N/A
Ours ($W_{k_z} \rightarrow W_m$ Only)	24.66	50.52	14.76	46.65
Ours (Mul$\rightarrow$Add Only)	25.61	50.92	15.50	47.13
Ours (SM only)	20.40	45.83	12.44	43.10
Ours (CM only)	23.24	46.15	14.71	45.32
Ours	**26.58**	**52.31**	**16.72**	**48.36**

Table 4. Trick ablation studies on FashionIQ and Shoes using ResNet-50. The score for FashionIQ is the average across all three sub-datasets.



(a) f_γ w/ Content keywords (b) f_γ w/ Style keywords

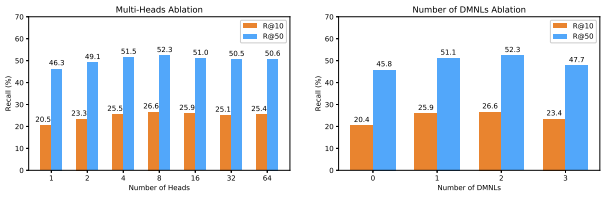


(c) f_β w/ Content keywords (d) f_β w/ Style keywords

Figure 5. t-SNE visualizations with the predictions of f_γ and f_β, given two different types of text inputs. For (a) and (c), we input modifier texts from that contain *content*-related keywords: `short, strap, long, print, sleeve`. For (b) and (d), we input text that contain *style*-related keywords: `red, black, blue, green, brown`.

mentation of MNL. We proposed two tricks used to alleviate the issue in 4.1.2. To further highlight the importance of these tricks, we conduct some ablations with the FashionIQ dataset [21]. As seen in the first three rows of Table 4, using either of the two tricks helps stabilize training and exhibits substantial improvement on Recall@10 over multi-level VAL [10]. Note that, when removing both tricks, we were unable to train the model to convergence due to the instability. These results imply that the gradient entanglement causes unstable training and that the proposed DMNL alleviates this issue significantly.

Behavior of SM To verify whether SM works as intended, we visualize the outputs of f_γ and f_β from Eq. (19). We select a set of style-related keywords (`red, black, blue, green, brown`) and randomly sample 300 modifier texts containing each of the words from the FashionIQ dataset. Similarly, we randomly sample from a set of content-related words (`short, strap, long, print, sleeve`) as well. Then, we visualize the outputs of f_γ and f_β from each set of modifier texts using t-SNE [63]. As depicted in Figure 5, SM produces more discriminative features for both f_γ and f_β given style-related modifiers (Figure 5(b,d)), as compared to those of content-related modifiers (Figure 5(a,c)).



(a) Multi-Head Ablation (b) DMNLs Stack Ablation

Figure 6. Effect of multi-heads and multi-layers on FashionIQ. The values indicate average scores across all three sub-datasets.

Effects of hyperparameters The proposed method has only two hyperparameters: the number of stacked DMNL blocks and the number of heads in each of them. We test the dependency of our compositor on these two hyperparameters by varying each of them separately. When varying the number of heads (Figure 6(a)), we observe that the performance gets saturated after 8 heads. With smaller number of heads, the model is not effective to capture the various semantics in the modifiers, and thus, results in weaker performances. Additionally, in Figure 6(a), we observe that using just one DMNL block is sufficient to outperform the state-of-the-art methods, and that using 3 or more DMNL blocks leads to suboptimal results due to over-parametrization.

7. Conclusion

In this work, we proposed a novel approach to image retrieval with text feedback. Our algorithm, CoSMo, is based on the idea of independently modulating the content and style of a reference image based on the given modifier text. Through our experiments and analysis, we demonstrated outstanding performance on multiple benchmarks and provided insights into the inner-workings of our image-text compositor. We hope that our work will influence future works to explore content and style modulation - not just in this specific task, but also in other tasks that require combining image and text features.

Acknowledgements This work was partly supported by Samsung Advanced Institute of Technology and Korean ICT R&D program of the MSIP/IITP grant [2017-0-01779, 2017-0-01780].

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019.

[2] Kenan E. Ak, Ashraf A. Kassim1, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *CVPR*, 2018.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[4] Arko Barman and Shishir K. Shah. A graph-based approach for making consensus-based decisions in image search and person re-identification. *IEEE TPAMI*, 2019.

[5] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[6] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*, 2017.

[7] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, 2019.

[8] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. 2020.

[9] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *CVPR*, 2020.

[10] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, 2020.

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[12] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018.

[13] Titir Dutta and Soma Biswas. s-sbir: Style augmented sketch based image retrieval. In *WACV*, 2020.

[14] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019.

[15] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In *CVPR*, 2019.

[16] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[17] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *ICCV*, 2019.

[18] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016.

[19] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Veronique Prinet, and Chunhong Pan. Progressive sparse local attention for video object detection. In *ICCV*, 2019.

[20] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *NeurIPS*, 2018.

[21] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019.

[22] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yinglong Wang, Jun Ma, and Mohan Kankanhalli. Attentive long short-term preference modeling for personalized product search. *ACM TOIS*, 2019.

[23] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[25] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[26] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *ICML*, 2018.

[27] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In *CVPR*, 2020.

[28] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[29] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *ICCV*, 2019.

[30] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

[31] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2019.

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[33] Justin Johnson, Alexandre AlahiLi, and Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[34] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[36] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *NeurIPS*. 2016.

[37] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *ICCV*, 2019.

[38] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.

[39] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[40] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *ICCV*, 2019.

[41] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016.

[42] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[43] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *ICLR*, 2017.

[44] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *ICLR*, 2020.

[45] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[47] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016.

[48] Jin Ma, Shanmin Pang, Bo Yang, Jihua Zhu, and Yaochen Li. Spatial-content image search in complex scenes. In *WACV*, 2020.

[49] Zhe Ma, Jianfeng Dong, Zhongzi Long, Yao Zhang, Yuan He, Hui Xue, and Shouling Ji. Fine-grained fashion similarity learning by attribute-specific embedding network. In *AAAI*, 2020.

[50] Long Mai, Hailin Jin, Zhe Lin, Chen Fang, Jonathan Brandt, and Feng Liu. Spatial-semantic image search by visual feature synthesis. In *CVPR*, 2017.

[51] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *NeurIPS*, 2018.

[52] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017.

[53] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 2016.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019.

[55] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[57] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM TOG*, 2016.

[58] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*. 2017.

[59] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, 2019.

[60] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, 2020.

[61] Rishab Sharma and Anirudha Vishvakarma. Retrieving similar e-commerce images using deep learning. *arXiv preprint arXiv:1901.03546*, 2019.

[62] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[63] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne, 2008.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[65] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[66] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019.

[67] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[68] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *ICCV*, 2019.

[69] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[70] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image-text matching. *IEEE TNNLS*, 2020.

[71] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *ECCV*, 2020.

[72] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016.

[73] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019.

[74] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. Context-aware attention network for image-text retrieval. In *CVPR*, 2020.

[75] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *CVPR*, 2017.

[76] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[77] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR*, 2020.