Dual Compositional Learning in Interactive Image Retrieval

Jongseok Kim[1], **Youngjae Yu**[1,2], **Hoeseong Kim**[1], **and Gunhee Kim**[1,2]
Seoul National University[1]
RippleAI[2], Seoul, Korea
{js.kim, yj.yu}@vision.snu.ac.kr, {hsgkim, gunhee}@snu.ac.kr

Abstract

We present an approach named *Dual Composition Network* (DCNet) for interactive image retrieval that searches for the best target image for a natural language query and a reference image. To accomplish this task, existing methods have focused on learning a composite representation of the reference image and the text query to be as close to the embedding of the target image as possible. We refer this approach as *Composition Network*. In this work, we propose to close the loop with *Correction Network* that models the difference between the reference and target image in the embedding space and matches it with the embedding of the text query. That is, we consider two cyclic directional mappings for triplets of (reference image, text query, target image) by using both Composition Network and Correction Network. We also propose a joint training loss that can further improve the robustness of multimodal representation learning. We evaluate the proposed model on three benchmark datasets for multimodal retrieval: Fashion-IQ, Shoes, and Fashion200K. Our experiments show that our DCNet achieves new state-of-the-art performance on all three datasets, and the addition of Correction Network consistently improves multiple existing methods that are solely based on Composition Network. Moreover, an ensemble of our model won the first place in Fashion-IQ 2020 challenge held in a CVPR 2020 workshop.

1 Introduction

Interactive conversational image search has risen as one of the next key technologies for search engines as it allows users to provide their search intent in a more efficient, intuitive, and engaging way. To implement such systems, it is a prerequisite to develop an algorithm that retrieves images based on previous search results and additional user feedback in a form of natural language queries, as shown in Figure 1. Not only does it have huge potential commercial values by enhancing the users' shopping experience, for example, but it is also an intriguing research problem to study multimodal embeddings and cross-modal retrieval.

In this work, we propose an approach named *Dual Composition Network* (DCNet) that can retrieve the best matching target image for a given reference image and a text query. Existing methods (Vo et al. 2019; Chen, Gong, and Bazzani 2020) have tackled this problem by training a network that can produce a composite feature combining the reference image and the text query and make it closer to the embedding of the corresponding target image. We call this model *Composition Network*. Going one step further, our key idea is to incorporate another network, which we call *Correction Network*, that models the difference feature between the target and reference image and pushes it closer to the embedding of the corresponding text query. In other words, as shown in Figure 1, our model can learn a more robust representation by considering cyclic directional mappings with no additional label, *i.e.* using both $I_r + T_q \rightarrow I_t$ and $I_t - I_r \rightarrow T_q$, where $I_{r/t}$ is the reference/target image and T_q is the text query.

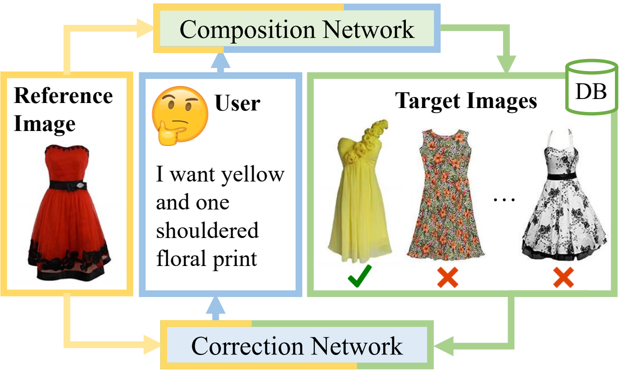

Figure 1: The key intuition of *DCNet*. Given a previously retrieved image (reference image), a user inputs how to update the search results in natural language (text query), and our goal is to search for the best matching image for them (target image). Our model considers both the forward (Composition Network) and inverse (Correction Network) pathways.

We summarize the contribution of this work as follows:

1. We propose *Correction Network* that drives the difference feature between the target and reference image to be similar to the corresponding text query embedding. Compared to existing models such as TIRG (Vo et al. 2019) and VAL (Chen, Gong, and Bazzani 2020) that rely only on Composition Network, our model complementarily models another directional mapping by Correction Network to learn more robust multimodal representations.

2. As a technical contribution, we propose a Fused Difference (FD) module to realize Correction Network and introduce a joint loss to synergistically train both networks. Our experiments show that they significantly improve the retrieval performance.

3. We achieve new state-of-the-art performance on three multimodal retrieval datasets: Fashion-IQ (Guo et al. 2019), Shoes (Guo et al. 2018), and Fashion200K (Han et al. 2017). Furthermore, we demonstrate that our key idea, the addition of Correction Network, is universally beneficial for existing methods that are solely based on Composition Network such as TIRG and VAL. Finally, our ensemble model won the first place in Fashion-IQ 2020 challenge in a CVPR 2020 workshop.

2 Related Work

Interactive Image Retrieval. Image retrieval is the task of retrieving the most relevant image for a given query. It has numerous applications across various domains such as generic images (Wang et al. 2014), faces (Schroff, Kalenichenko, and Philbin 2015), persons (Zheng et al. 2015), birds (Forbes et al. 2019), fashion (Liu et al. 2016; Ge et al. 2019), and landmarks (Weyand et al. 2020), to name a few. In particular, the use of text query has been widely studied in the field of computer vision and language processing including the retrieval of generic images (Karpathy and Li 2015; Klein et al. 2015; Wang, Li, and Lazebnik 2016) and person images (Li et al. 2017). We explore the type of image retrieval where both a reference image and a text query are provided.

Interactive image retrieval goes one step further by utilizing user feedback to refine the search results. Feedback is provided in a multitude of different formats, including attribute labels (Isola, Lim, and Adelson 2015; Zhao et al. 2017; Han et al. 2017; Ak et al. 2018), relative attributes (*e.g. longer* or *brighter*) (Parikh and Grauman 2011; Kovashka, Parikh, and Grauman 2012; Kovashka and Grauman 2013), synthetic sentences (Vo et al. 2019), and human annotated sentences (Guo et al. 2018; Tan et al. 2019; Guo et al. 2019; Yu, Shen, and Jin 2020). In this work, we mainly deal with natural sentence queries.

Recently, many deep learning methods have been proposed for interactive image retrieval. (Guo et al. 2019) apply additive attention mechanism, (Vo et al. 2019) propose Text Image Residual Gating (TIRG) to compose the reference image and text query with a gating function and residual connection, (Hosseinzadeh and Wang 2020) decompose an image into a set of local semantic entities, and (Chen, Gong, and Bazzani 2020) propose VAL that utilizes intermediate features of images to which text features attend for composition. (Chen and Bazzani 2020) enhance image representation by utilizing extra textual information that directly describes the image. Most existing methods are confined to devising a better feature composition between images and text. On the other hand, our approach extends this idea by additionally incorporating the difference between the target and reference image.

Difference Detection between Images. Whereas learning the pixel-level difference and/or flows of consecutive images have been extensively studied (Stent et al. 2016; Khan et al. 2017), only a few attempts have been made to model the difference between two image embeddings. (Guo et al. 2018) use feature concatenation or feature fusion using linear layers or convolution layers. (Guo et al. 2019) create a representation of the difference by simply subtracting two visual embeddings. (Park, Darrell, and Rohrbach 2019) improve the representation of the difference with dual attention mechanism.

Multi-task Objectives for Multimodal Tasks. Multi-task joint training has often been employed to improve the performance by training relevant tasks together. In image captioning, (Liu et al. 2018) incorporate image retrieval to generate more diverse and richer sentences and (Liu et al. 2020) add a text retrieval module to improve the quality of generated captions. (Qiao et al. 2019; Joseph et al. 2019) jointly train an image generation model with a captioning model for better text-to-image generation. (Shah et al. 2019) train a visual question answering module with question generation task to be more robust to linguistic variations. (Xu et al. 2015) build a joint embedding model that tackles language generation as well as video retrieval and language retrieval. Our approach improves interactive image retrieval by training both Composition and Correction Network with multi-task training objectives.

3 Approach

Our goal is to retrieve the best target image from the image database (*i.e.* $I_t \in \mathcal{I}$) for a given text query T_q with respect to the reference image I_r. Figure 2 outlines the overall architecture of our *Dual Composition Network* (DCNet), which consists of two networks.

First, *Composition Network* learns a composition feature that combines the reference image and the text query so that it matches well with that of the correct target image. We advance the existing composition networks by introducing *experts* that generate features by attending to different parts of the images and text. Second, *Correction Network* represents the difference between the reference and target image in the same embedding space with the query text. Finally, we introduce a joint loss that facilitates the training of both networks in concert with each other to yield better performance.

3.1 Experts

Inspired by recent works that utilize multiple embeddings for visual representation (Hosseinzadeh and Wang 2020; Chen, Gong, and Bazzani 2020), we divide an image and a sentence into a set of localized components and assign a representation module denoted *expert* to each of them. That is, we represent both images and text as a combination of multiple features encoded by multiple experts that are specialized for different parts.

Image Experts. We encode each image with E number of experts, which are divided into two types: experts for different intermediate layers and for different spatial locations.

As for the layer experts $\mathbf{x}_{lay}$, we first extract features from L intermediate layers of the backbone CNN:

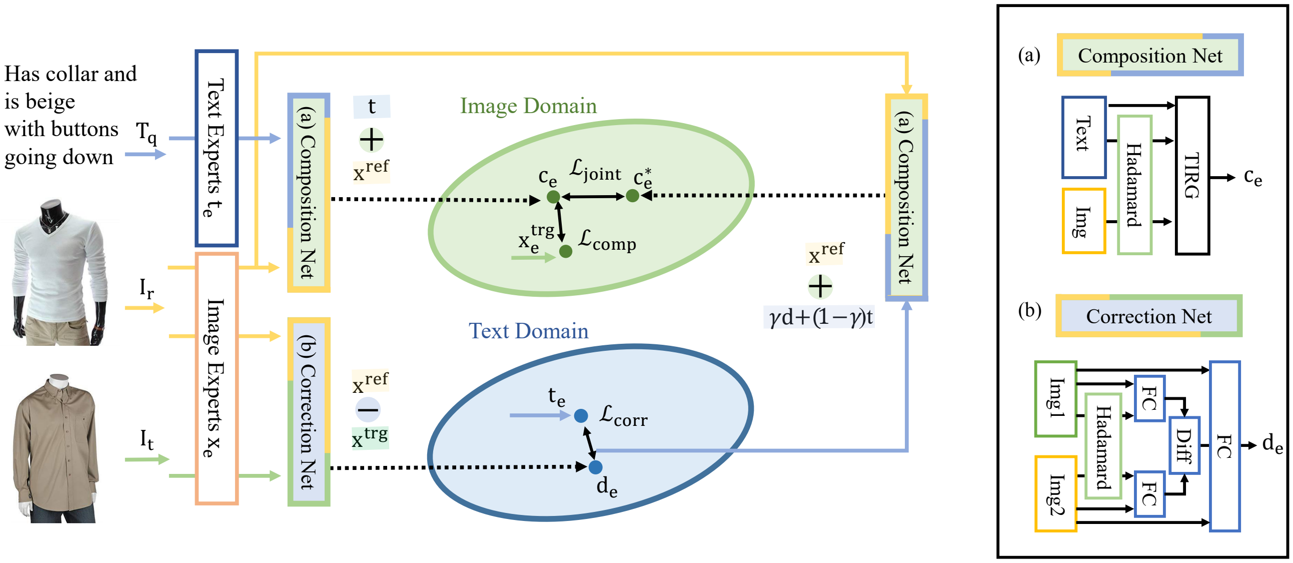

Figure 2: Overall architecture of our proposed DCNet. For a triplet of (I_t, I_t, T_q), Image and Text Experts obtain image and text representation (section 3.1). (a) Composition Network computes the composition feature between reference image I_r and text query T_q and then matches it with the target image representation I_t ($\mathcal{L}_{comp}$). (b) Correction Network computes the difference feature between the reference and target image then matches it with the text query representation ($\mathcal{L}_{corr}$) (section 3.2). The whole model is jointly trained with additional joint loss ($\mathcal{L}_{joint}$) (section 3.3).

$\{\phi^1, \phi^2, \cdots, \phi^{L-1}, \hat{\phi}\} = f_{\text{CNN}}(I)$. For example, we use ResNet50 for f_{CNN} with $L = 2$, and choose the outputs from layer3 and the last layer for ϕ^1 and $\hat{\phi}$. Then, each layer expert encodes the intermediate output by applying global average pooling followed by an FC layer to each ϕ^i ($\phi^L = \hat{\phi}$):

$$\mathbf{x}_{lay,i} = \text{FC}(\text{Avg}(\phi^i)) \text{ for } i = 1, \cdots, L. \tag{1}$$

For interactive image retrieval, a user may want to change only a part of an image such as collar or sleeve of a shirt. To better focus on specific parts of the image, we add spatial experts $\mathbf{x}_{spat}$, each of which is obtained from a particular location of the last layer $\hat{\phi}$ with an FC layer:

$$\mathbf{x}_{spat,i} = \text{FC}(\hat{\phi}[w_i, h_i]) \text{ for } i = 1, \cdots, P. \tag{2}$$

(w_i, h_i) denotes the location of $\hat{\phi}$; for example, $(w_i, h_i) \in \{(3,1),(1,3),(3,3),(5,3),(3,5)\}$ for ResNet50 with $\hat{\phi} \in \mathbb{R}^{7\times7\times2048}$.

Finally, for a total of $E = L+P$ image experts, the image representation $\mathbf{x}$ is obtained as $\mathbf{x} = \mathbf{x}_{lay} \cup \mathbf{x}_{spat}$. Note that each expert has distinct FC parameters.

Text Experts. Given a text query with l words, we first embed each word using GloVe (Pennington, Socher, and Manning 2014) as $w^* = [w_1^*, ..., w_l^*] \in \mathbb{R}^{l\times300}$. We then obtain the query embedding $w \in \mathbb{R}^{l\times D}$ by

$$w = \text{FC}([\text{Conv1d}(w^*); w^*]), \tag{3}$$

where $[;]$ denotes concatenation.

Preferably, each image expert should attend to different words in the text. In the example of Figure 1, experts for the whole image should attend to *yellow* or *less flowy*, while experts focusing on the top part of the image should attend to *shouldered floral print*. To implement this idea, we obtain the query embedding $t_e^* \in \mathbb{R}^D$ for each expert e as

$$t_e^* = \sum_l \alpha_{e,l} w_l^*, \quad \alpha_e = \underset{l}{\text{softmax}}(\text{FC}(\text{FC}(m_e \odot w))), \tag{4}$$

where $m_e \in \mathbb{R}^D$ is a randomly initialized vector and $\odot$ is Hadamard product. Finally, we obtain text representation $\mathbf{t}$ by applying an expert specific FC layer to t_e^* as $\mathbf{t} = \cup_{e=1}^{E} \text{FC}(t_e^*)$. Note that the number of text experts is also E as each image expert corresponds to one text expert.

3.2 The Dual Composition Network

Composition Network. Composition Network learns to combine the features of the reference image and the text query to be as similar as possible to the feature of the correct target image. As shown in Figure 2(a), we adopt a variant of Text Image Residual Gating (TIRG) (Vo et al. 2019), whose principal idea is to perturb the gated feature with a residual connection to obtain the composite feature:

$$\bar{t}_e = [t_e; \text{Fusion}(x_e^{ref}, t_e)], \tag{5}$$

$$c_e = w_g f_{gate}(x_e^{ref}, \bar{t}_e) + w_r f_{res}(x_e^{ref}, \bar{t}_e), \tag{6}$$

where w_g and w_r are learnable parameters. We choose Hadamard product as the fusion function. The gating and residual connections are computed by

$$f_{gate}(x_e^{ref}, \bar{t}_e) = \sigma(\text{FC}([x_e^{ref}; \bar{t}_e])) \odot x_e^{ref} \tag{7}$$

$$f_{res}(x_e^{ref}, \bar{t}_e) = \text{FC}(\text{FC}([x_e^{ref}; \bar{t}_e])) \tag{8}$$

where σ is the sigmoid function.

Finally, the matching score of Composition Network s^r for a triplet is evaluated as

$$s^r(\mathbf{x}^{ref}, \mathbf{t}, \mathbf{x}^{trg}) = \sum_e c_e \cdot x_e^{trg}. \tag{9}$$

Correction Network. Correction Network helps the retrieval of the correct target image by representing the difference between the reference and target image and pushing it closer to the text query embedding. We name the following implementation as *Fused Difference* (FD) module since it can be used as a part of any multimodal embedding.

As shown in Figure 2(b), we first represent the difference between the reference and target image as

$$\bar{x}_e^{diff} = \bar{x}_e^{trg} - \bar{x}_e^{ref} \quad \text{where} \tag{10}$$

$$\bar{x}_e^{trg/ref} = \text{FC}^{trg/ref}([x_e^{trg} \odot x_e^{ref}; x_e^{trg/ref}]). \tag{11}$$

We then compute the difference feature as

$$d_e = \text{FC}([x_e^{ref}; x_e^{trg}; \bar{x}_e^{diff}]). \tag{12}$$

The intuition is that we concatenate the difference feature with the original reference and target features and let the network learn to identify which feature dimensions are salient in the difference.

Finally, we calculate the triplet score of Correction Network by

$$s^c(\mathbf{x}^{ref}, \mathbf{x}^{trg}, \mathbf{t}) = \sum_e d_e \cdot t_e. \tag{13}$$

3.3 Joint Training and Inference

We jointly train all of the components of DCNet, including Composition and Correction Networks with image and text experts. For a minibatch of size B that consists of ground truth triplets $\{\mathbf{x}_i^{ref}, \mathbf{t}_i, \mathbf{x}_i^{trg}\}_{i=1}^B$, the loss functions for the two networks are defined as the cross-entropy loss:

$$\mathcal{L}_{comp} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s^r(\mathbf{x}_i^{ref}, \mathbf{t}_i, \mathbf{x}_i^{trg}))}{\sum_{j=1}^{B} \exp(s^r(\mathbf{x}_i^{ref}, \mathbf{t}_i, \mathbf{x}_j^{trg}))}, \tag{14}$$

$$\mathcal{L}_{corr} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s^c(\mathbf{x}_i^{ref}, \mathbf{x}_i^{trg}, \mathbf{t}_i))}{\sum_{j=1}^{B} \exp(s^c(\mathbf{x}_i^{ref}, \mathbf{x}_j^{trg}, \mathbf{t}_i))}. \tag{15}$$

where s^r and s^c are the scores obtained from Composition and Correction Network (Eq. (9) and Eq. (13)), respectively.

In addition, we introduce another joint loss $\mathcal{L}_{joint}$ that connects the two networks:

$$\mathcal{L}_{joint} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s^j(\mathbf{x}_i^{ref}, \mathbf{x}_i^{trg}, \mathbf{t}_i))}{\sum_{j=1}^{B} \exp(s^j(\mathbf{x}_i^{ref}, \mathbf{x}_j^{trg}, \mathbf{t}_i))}, \tag{16}$$

where

$$s^j = \sum_e c_e \cdot \text{COMP}(x_e^{ref}, (\gamma d_e + (1-\gamma)t_e)). \tag{17}$$

COMP denotes the composition network, and $\gamma = 0.5$ is the hyperparameter for perturbation. The intuition behind score s^j emanates from the idea that the joint loss should ultimately serve as a bond between the two seemingly opposite Networks. The resultant feature d_e should ideally be adequate as a replacement for t_e, as they both model the difference. Therefore, replacing x_e^{trg} in Eq. (9) with COMP$(x_e^{ref}, \gamma d_e + (1-\gamma)t_e)$ should have the same effect as Eq. (9) and connect the two Networks. In our experiments, the addition of joint loss $\mathcal{L}_{joint}$ helps both networks more to be robust to text variations.

The final loss function for joint training is the weighted sum of these three losses:

$$\mathcal{L} = \mathcal{L}_{comp} + \mathcal{L}_{corr} + \lambda\mathcal{L}_{joint}, \tag{18}$$

where λ is a hyperparameter (*e.g.* $\lambda = 0.5$ for experiments).

Inference. At test time, the final score s for a triplet is calculated as $s = s^r + s^c$.

4 Experiments

We assess the performance of our approach on three benchmark datasets, including Fashion-IQ (Guo et al. 2019), Shoes (Guo et al. 2018) and Fashion200K (Han et al. 2017). The natural sentence queries are collected by humans in the first two datasets but synthesized in the last dataset. Following previous works on these datasets, we report the recall at rank k (Recall@k) as the evaluation metric, specifically Recall@10 (R@10) and Recall@50 (R@50).

Datasets. (1) Fashion-IQ (Guo et al. 2019) is an interactive image retrieval dataset that contains 30,134 triplets from 77,683 fashion images of three categories (*i.e.* Dress, Shirt and Tops&Tees) crawled from Amazon.com.

(2) Shoes (Guo et al. 2018) is a dataset based on images crawled from like.com (Berg, Berg, and Shih 2010). For interactive image retrieval, natural language query sentences are additionally obtained from human annotators. Following (Chen, Gong, and Bazzani 2020), we use 10K images for training and 4,658 images for evaluation.

(3) Fashion200K (Han et al. 2017) contains about 200K fashion images. Following (Vo et al. 2019), we pair two images that have only one word difference in their descriptions as reference and target images to synthesize query sentences. As done in (Vo et al. 2019), we use about 172K triplets for training and 33,480 triplets for evaluation.

Implementation. For a fair comparison with VAL (Chen, Gong, and Bazzani 2020), we select ResNet50 (He et al. 2016) pretrained on ImageNet (Wang et al. 2014) as our backbone encoder. We also employ random cropping and horizontal flipping for image augmentation. We extract GloVe from text using the `en_vectors_web_lg` model of `spaCy`. We set the hidden dimension to 1024 and use ReLU activation for every FC layer with dropout (Srivastava et al. 2014) of rate 0.2. We apply L_2 normalization to image and text embeddings. Each training batch contains $B = 32$ triplets of (reference image, text query, target image) and is shuffled at the beginning of every training epoch. We use Adam (Kingma and Ba 2015) optimizer with a learning rate of 1×10^{-4} and an exponential decay of 0.95 at every epoch. All models are implemented with PyTorch.

Category	Dress		Shirt		Toptee		Total
Metric	R@10	R@50	R@10	R@50	R@10	R@50	Avg
Side+Gating (Guo et al. 2019)	11.24	32.39	13.73	37.03	13.52	34.73	23.77
FiLM (Perez et al. 2018)‡	14.23	33.34	15.04	34.09	17.30	37.68	25.28
TIRG (Vo et al. 2019)‡	14.87	34.66	18.26	37.89	19.08	39.62	27.40
Relationship (Santoro et al. 2017)‡	15.44	38.08	18.33	38.63	21.10	44.77	29.39
VAL ($\mathcal{L}_{vv}$) (Chen, Gong, and Bazzani 2020)‡	21.12	42.19	21.03	43.44	25.64	49.49	33.82
VAL (GloVe) (Chen, Gong, and Bazzani 2020)‡	22.53	44.00	22.38	44.15	27.53	51.68	35.38
CurlingNet (Yu et al. 2020)	26.15	53.24	21.45	44.56	30.12	55.23	38.45
DCNet	**28.95**	**56.07**	**23.95**	**47.30**	**30.44**	**58.29**	**40.83**

Table 1: Comparison between our DCNet and baselines on the Fashion-IQ validation set. ‡ denotes the results cited from (Chen, Gong, and Bazzani 2020).

Participants	AvgR@(10,50)
cuberick	0.33
emd	0.39
skywalker (Li et al. 2019)	0.44
shuan (Yu et al. 2020)	0.47
superraptors	0.49
zyday	0.43
stellastra	0.48
ganfu.wb	0.50
tianxi.tl	0.50
Ours	**0.52**

Table 2: Final results on the Test Phase of Fashion-IQ 2020 Challenge. We also report the results of last year's participants (Fashion-IQ 2019) in the upper part. Note that the use of attribute metadata for the test split is prohibited this year, which was a crucial reason of strong performance of last year's methods. Our method won the first place in the challenge.

4.1 Results on Fashion-IQ

Table 1 shows the quantitative results on Fashion-IQ validation set. Our DCNet approach achieves the new state-of-the-art performance over all three categories of the dataset.

Challenge. To demonstrate the practical competence of our method, we participated in Fashion-IQ 2020 challenge organized in a CVPR 2020 workshop[1]. Table 2 summarizes the official leaderboard, where our ensembled method won the first place. For a better representation of fashion related features, we use a backbone image encoder pre-trained on Deepfashion (Liu et al. 2016) attribute prediction task. We ensemble the following variants of our method:

(1) Image encoders: ResNet50/152, DenseNet169 (Huang et al. 2017).
(2) Text embeddings: BERT (Devlin et al. 2019).
(3) Text encoders: FC, and average pooling.
(4) Image experts: 3 layer experts (*i.e.* $L = 3$), no spatial expert (*i.e.* $P = 0$)
(5) Hidden dimension sizes (512 and 1024), batch sizes (16 and 32) and 5 different random seeds.

DCNet Variants	AvgR@10	AvgR@50	Avg
Separate (Comp)	26.14	51.82	38.98
Separate (Corr)	26.44	52.82	39.63
Experts (Comp)	26.41	52.68	39.55
Experts (Corr)	26.75	52.81	39.78
Joint (Comp)	26.61	52.85	**39.73**
Joint (Corr)	27.21	53.41	**40.31**

Table 3: Ablation results of our joint training procedure on Fashion-IQ validation set. (Comp) and (Corr) denote the results of using only Composition and Correction Network at test time, respectively. (Separate) is the separate training for the two networks, (Experts) jointly train only image and text experts without the joint loss $\mathcal{L}_{joint}$, and (Joint) follows our full training procedure.

Methods	AvgR@10	AvgR@50	Avg
Concat	20.21	45.33	32.77
+ Joint	20.96	46.12	33.54
TIRG	20.72	45.88	33.31
+ Joint	21.73	46.01	33.87
VAL	24.37	49.90	37.13
+ Joint	25.35	51.02	38.19

Table 4: Results of joint training with Correction Network for three existing methods on the Fashion-IQ validation split.

Ablation for Joint Training. Table 3 shows the results of ablation studies on our joint training procedure by varying the level of joint training: (1) Separate: Composition and Correction Networks are trained separately, (2) Experts: only image and text experts are jointly trained without the joint loss $\mathcal{L}_{joint}$, and (3) Joint: our joint loss is used. (Comp) and (Corr) respectively indicate that only Composition and Correction Network is used for retrieval at test time. The results show that the both networks gradually improve in performance as more components are jointly trained.

Effects of Correction Network. Our key novelty lies in

[1] https://sites.google.com/view/cvcreative2020/fashion-iq

Methods	AvgR@10	AvgR@50	Avg
Concat	19.32	42.79	31.06
Diff	21.42	45.83	33.63
Diff-expert	24.75	50.59	37.67
TIRG	10.85	30.14	20.49
DUDA	22.21	48.25	35.23
Ours-singleText	26.34	51.18	38.75
Ours	**26.44**	**52.82**	**39.63**

Table 5: Comparison between the difference representations for images on the Fashion-IQ validation split. Ours is the proposed Fused Difference (FD) module and -singleText indicates that only a single text expert is used. Diff-expert indicates that multiple experts are used for image representation.

Methods	R@10	R@50
FiLM†	38.89	68.30
TIRG†	45.45	69.39
Relationship†	45.10	71.45
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$)†	49.12	73.53
VAL (GloVe)†	51.52	75.83
Separate (Comp)	51.24	78.04
Separate (Corr)	51.70	78.75
Joint (Comp)	51.88	78.53
Joint (Corr)	53.24	78.67
DCNet	**53.82**	**79.33**

Table 6: Results on the Shoes validation split. Baseline results are shown in top, while those of variants of our DCNet are in bottom. Please refer to the caption of Table 3 for the variants. †: results from (Chen, Gong, and Bazzani 2020).

Methods	R@10	R@50
FiLM†	39.5	61.9
Relationship†	40.5	62.4
TIRG†	42.5	63.8
VAL ($\mathcal{L}_{vv}$)‡	45.75	66.10
TIRG*	45.62	67.24
Separate (Corr)	44.18	63.41
Joint (TIRG)	46.31	67.41
Joint (Corr)	45.96	66.62
DCNet	**46.89**	**67.56**

Table 7: Results on Fashion200K. †: results from (Vo et al. 2019), ‡: results from the execution of the official code[2], *: re-implementation using MobileNetV1 (Howard et al. 2017) as the backbone.

the introduction of Correction Network to close the loop with Composition Network for a synergetic performance improvement. Thus, we validate that the joint training with Correction Network universally benefits existing methods that take Composition Network only approach. We test with three types of existing methods for composition networks: Concat (Guo et al. 2018, 2019), TIRG (Vo et al. 2019) and VAL (Chen, Gong, and Bazzani 2020). Concat indicates the simple concatenation between the embeddings of the reference image and the text query, followed by a single FC layer. For image embedding of Concat and TIRG, we use a globally pooled feature from the last convolution layer. For that of VAL, we use the output from three intermediate layers (layer 2, 3, and 4) following the original paper. To match the feature dimension, we apply global pooling before Correction Network.

Table 4 compares the results when our design of Composition Network is replaced with three conventional methods. Correction Network improves R@10 and R@50 of Concat by 0.75, 0.79 on average, TIRG by 1.01, 0.13, and VAL by 0.98 and 1.12, respectively. These results indicate that our Correction Network is universally beneficial for existing state-of-the-art methods.

Difference Representation. To validate our design of the Fused Difference module in Correction Network, we test several techniques to obtain the difference representation between the reference and target images. Table 5 summarizes the results. Concat simply concatenates the reference and target image features and applies an FC layer, and DIFF additionally concatenates the subtraction of two features before the FC layer. We also compare with some state-of-the-art fusion models including TIRG and DUDA (Park, Darrell, and Rohrbach 2019). For a fair comparison, we use our image expert representation in all techniques ($\mathbf{x} \in \mathbb{R}^{E \times D}$). For TIRG, Fusion, Concat and Diff, we average the image representation ($\mathrm{Avg}(\mathbf{x}) \in \mathbb{R}^{D}$). In the table, Diff-expert indicates Diff but with multiple image experts and Ours-singleText is our model but with only a single text expert as other baselines use only one as well. The results show that our FD module outperforms other methods. Comparison between Diff and TIRG shows that the summation-based TIRG is not capable of representing the difference between features. Furthermore, Diff surpasses Concat by 2.57 and DUDA (Park, Darrell, and Rohrbach 2019) outperforms Diff by 1.6 in average recall. Our FD module with a single text expert surpasses DUDA by 3.52 and Diff-expert by 1.08, which indicates the proposed FD module can generate a better difference representation.

4.2 Results on Shoes

Table 6 summarizes the quantitative results on Shoes. Even when using only our Composition Network, denoted Separate (Comp), we can achieve comparable performance with the SOTA VAL (Chen, Gong, and Bazzani 2020) in R@10 and outperforms all conventional methods in R@50. The addition of Correction Network significantly improves the performance in both metrics. Note that in contrast to VAL, the scores of our methods are attained without pretraining the networks with additional descriptive texts (*i.e.* captioning sentences).

4.3 Results on Fashion200K

Table 7 shows the results on Fashion200K. While ResNet-50 is used for all methods in previous experiments, we here test with two CNN backbones: ResNet-18 for top three methods

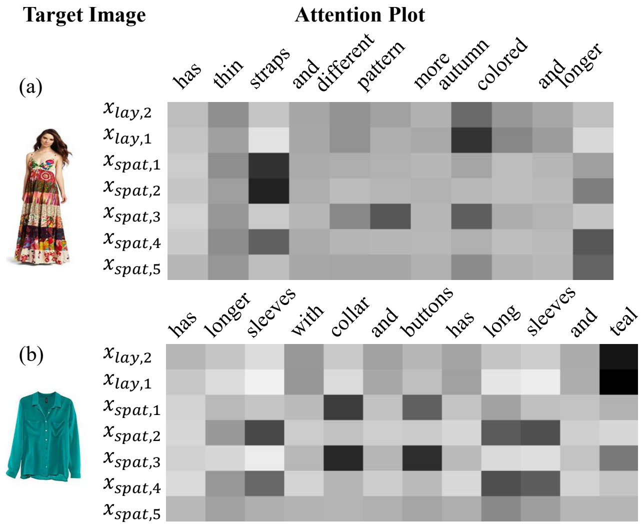

Figure 3: Visualization of attention weights $\alpha_{e,l}$ of text experts for the retrieved samples. $x_{lay,i}$ denote layer experts and $x_{spat,i}$ denote spatial experts corresponding to the up, left, center, right, and bottom section.

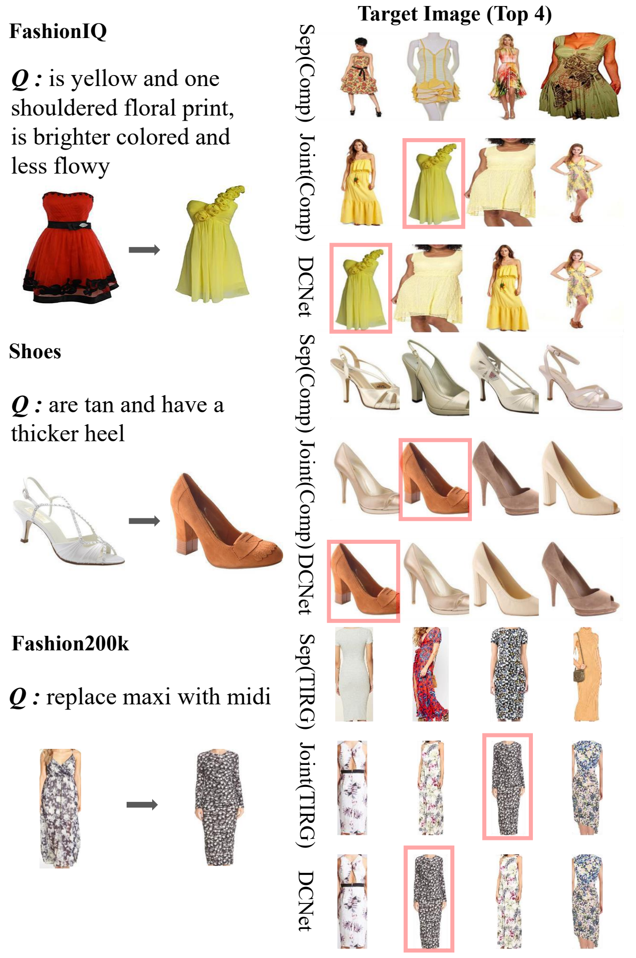
FashionIQ
Q : is yellow and one shouldered floral print, is brighter colored and less flowy
Shoes
Q : are tan and have a thicker heel
Fashion200k
Q : replace maxi with midi
Target Image (Top 4)
Sep(Comp)
Joint(Comp)
DCNet
Sep(Comp)
Joint(Comp)
DCNet
Sep(TIRG)
Joint(TIRG)
DCNet

Figure 4: Examples of the variants of our DCNet. Please refer to the caption of Table 3 for the variants.

in the table and MobileNetV1 (Howard et al. 2017) for bottom six methods, since these two have been used in previous works. For image and text embedding, we follow TIRG (Vo et al. 2019) that uses global pooling of the last CNN layer for images and LSTMs for text. The results demonstrate that our dual network approach is also powerful for a dataset with synthesized text queries that are far simpler than human annotated natural language sentences. For example, our Correction Network combined with TIRG, denoted DCNet, improves TIRG by 1.27 and 0.32 on R@10 and R@50, respectively.

4.4 Qualitative Results

Figure 3 visualizes the attention weights $\alpha_{e,l}$ of experts for each word, where $x_{lay,i}$ and $x_{spat,i}$ each denotes layer and spatial experts. Specifically, $x_{spat,1}, \cdots, x_{spat,5}$ focus on the up, left, center, right and bottom part of the image, respectively. We can observe that the layer experts attend to color-related words such as "autumn colored" and "teal", while the spatial experts attend to words describing specific parts of clothes such as "straps" (up, left, and right), "longer" (bottom), "buttons" and "collars" (up and center), and "sleeves" (left and right). This confirms that the proposed module matches each word to the appropriate image expert as intended.

Figure 4 qualitatively assesses the performance of our model while varying the joint training procedure, and Figure 5 qualitatively compares our DCNet with TIRG (Vo et al. 2019) and VAL (Chen, Gong, and Bazzani 2020). In the left, we present the correct triplets of (query, reference, target) and compare the top-4 ranked images from different models. Thanks to double checking by Correction Network, our joint method can better find the best matching target images that satisfy all conditions in the queries than all other models compared. Figure 6 further illustrates the retrieval results of our method for some test examples on three datasets. We display reference images and text queries in the left and 4 highest scored images in the right. While maintaining the style of the reference images, our model correctly assigns high scores to the samples that adequately reflect the description of user queries.

5 Conclusion

In this work, we proposed Dual Composition Network (DCNet) for interactive image retrieval with a natural language query. The two key components of the model, Composition and Correction Networks, were indeed synergetic to improve the performance of text-based image retrieval; as a result, our method achieved new state-of-the-art performance on Fashion-IQ, Shoes and Fashion200K, and won the first place in Fashion-IQ 2020 challenge. Interestingly, our idea of closing the loop with Correction Network was general enough to improve the performance of existing Composition Network only methods.

Moving forward, it is an interesting future work to ex-

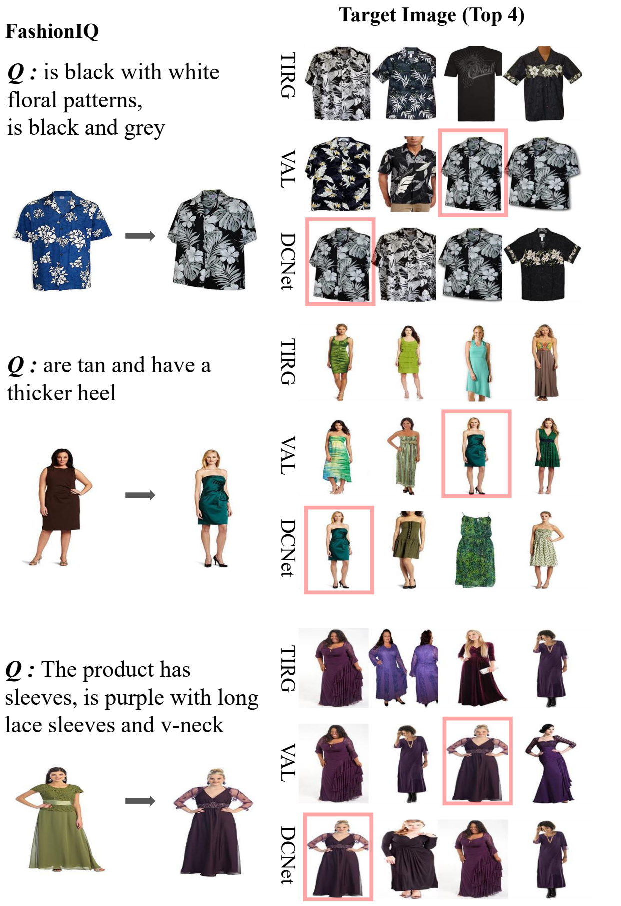
FashionIQ
Target Image (Top 4)
Q : is black with white floral patterns, is black and grey
TIRG
VAL
DCNet
Q : are tan and have a thicker heel
TIRG
VAL
DCNet
Q : The product has sleeves, is purple with long lace sleeves and v-neck
TIRG
VAL
DCNet

Figure 5: Qualitative comparison with TIRG (Vo et al. 2019) and VAL (Chen, Gong, and Bazzani 2020). We compare the results sampled from Fashion-IQ validation split.

pand the applicability of DCNet; we can modify our model to solve multi-hop interactive image search tasks, or we can tackle relative caption or dialogue generation tasks for multimodal conversational systems.

Acknowledgments

We thank SNUVL lab members for helpful comments. This research was supported by AIR Lab (AI Research Lab) in Hyundai Motor Company through HMC-SNUAI Consortium Fund and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01082, SW StarLab and No. 2017-0-01772, Video Turing Test). Jongseok Kim was supported by Hyundai Motor Chung Mong-Koo Foundation. Gunhee Kim is the corresponding author.

References

Ak, K. E.; Kassim, A. A.; Lim, J. H.; and Tham, J. Y. 2018. Learning Attribute Representations with Localization for Flexible Fashion Search. In *CVPR*.

Berg, T. L.; Berg, A. C.; and Shih, J. 2010. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *ECCV*.

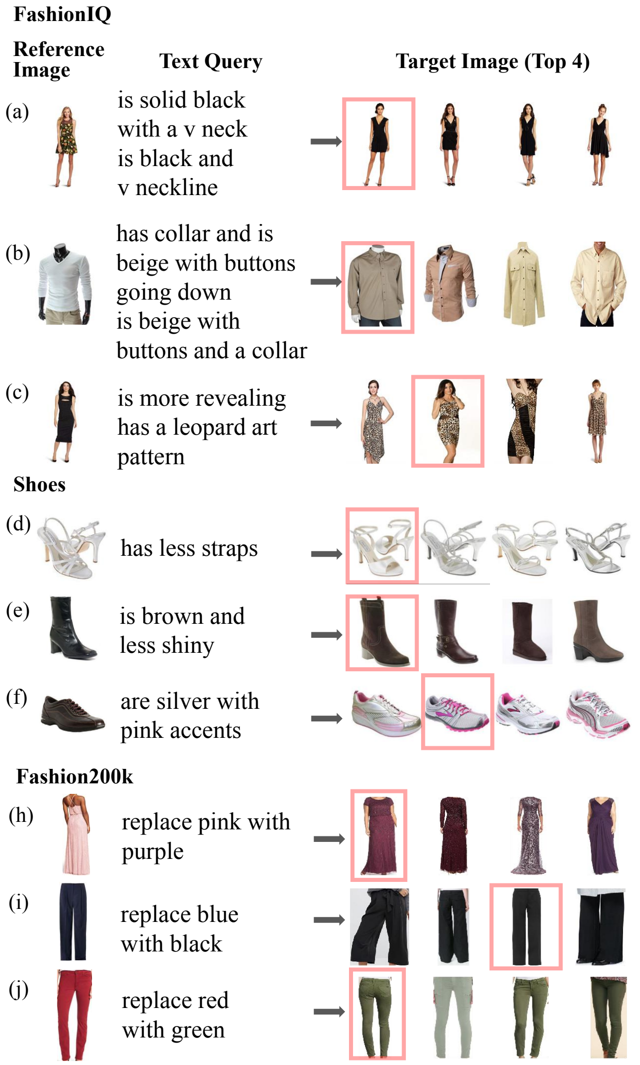
FashionIQ
Reference Image
Text Query
Target Image (Top 4)
(a)
is solid black with a v neck is black and v neckline
(b)
has collar and is beige with buttons going down is beige with buttons and a collar
(c)
is more revealing has a leopard art pattern
Shoes
(d)
has less straps
(e)
is brown and less shiny
(f)
are silver with pink accents
Fashion200k
(h)
replace pink with purple
(i)
replace blue with black
(j)
replace red with green

Figure 6: Some retrieval examples of our DCNet for the test samples of Fashion-IQ, Shoes and Fashion200K.

Chen, Y.; and Bazzani, L. 2020. Learning Joint Visual Semantic Matching Embeddings for Language-guided Retrieval. In *ECCV*.

Chen, Y.; Gong, S.; and Bazzani, L. 2020. Image Search with Text Feedback by Visiolinguistic Attention Learning. In *CVPR*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Forbes, M.; Kaeser-Chen, C.; Sharma, P.; and Belongie, S. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. In *EMNLP*.

Ge, Y.; Zhang, R.; Wu, L.; Wang, X.; Tang, X.; and Luo, P. 2019. A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. In *CVPR*.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; and Feris, R. 2018. Dialog-based Interactive Image Retrieval. In *NeurIPS*.

Guo, X.; Wu, H.; Gao, Y.; Rennie, S.; and Feris, R. 2019. The Fashion IQ Dataset: Retrieving Images by Combining Side Information and Relative Natural Language Feedback. In *ICCV*.

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic Spatially-aware Fashion Concept Discovery. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Hosseinzadeh, M.; and Wang, Y. 2020. Composed Query Image Retrieval Using Locally Bounded Features. In *CVPR*.

Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* .

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *CVPR*.

Isola, P.; Lim, J.; and Adelson, E. 2015. Discovering states and transformations in image collections. In *CVPR*.

Joseph, K. J.; Pal, A.; Rajanala, S.; and Balasubramanian, V. N. 2019. C4Synth: Cross-Caption Cycle-Consistent Text-to-Image Synthesis. In *WACV*.

Karpathy, A.; and Li, F. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*.

Khan, S. H.; He, X.; Porikli, F. M.; Bennamoun, M.; Sohel, F.; and Togneri, R. 2017. Learning Deep Structured Network for Weakly Supervised Change Detection. In *IJCAI*.

Kingma, D.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Klein, B.; Lev, G.; Sadeh, G.; and Wolf, L. 2015. Associating Neural Word Embeddings with Deep Image Representations using Fisher Vectors. In *CVPR*.

Kovashka, A.; and Grauman, K. 2013. Attribute Pivots for Guiding Relevance Feedback in Image Search. In *ICCV*.

Kovashka, A.; Parikh, D.; and Grauman, K. 2012. WhittleSearch: Image Search with Relative Attribute Feedback. In *CVPR*.

Li, J.; whan Lee, J.; sang Song, W.; young Shin, K.; and hyun Go, B. 2019. Designovel's System Description for Fashion-IQ Challenge 2019. *arXiv* .

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person Search with Natural Language Description. In *CVPR*.

Liu, J.; Wang, K.; Xu, C.; Zhao, Z.; Xu, R.; Shen, Y.; and Yang, M. 2020. Interactive Dual Generative Adversarial Networks for Image Captioning. In *AAAI*.

Liu, X.; Li, H.; Shao, J.; Chen, D.; and Wang, X. 2018. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. In *ECCV*.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *CVPR*.

Parikh, D.; and Grauman, K. 2011. Relative attributes. In *ICCV*.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust Change Captioning. In *ICCV*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual Reasoning with a General Conditioning Layer. In *AAAI*.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. MirrorGAN: Learning Text-To-Image Generation by Redescription. In *CVPR*.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *NIPS*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *CVPR*.

Shah, M.; Chen, X.; Rohrbach, M.; and Parikh, D. 2019. Cycle-Consistency for Robust Visual Question Answering. In *CVPR*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *JMLR* .

Stent, S.; Gherardi, R.; Stenger, B.; and Cipolla, R. 2016. Precise Deterministic Change Detection for Smooth Surfaces. In *WACV*.

Tan, F.; Cascante-Bonilla, P.; Guo, X.; Wu, H.; Feng, S.; and Ordonez, V. 2019. Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries. In *NeurIPS*.

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 *CVPR*.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning Fine-Grained Image Similarity with Deep Ranking. In *CVPR*.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *CVPR*.

Weyand, T.; Araujo, A.; Cao, B.; and Sim, J. 2020. Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *CVPR*.

Xu, R.; Xiong, C.; Chen, W.; and Corso, J. J. 2015. Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework. In *AAAI*.

Yu, T.; Shen, Y.; and Jin, H. 2020. Towards Hands-Free Visual Dialog Interactive Recommendation. In *AAAI*.

Yu, Y.; Lee, S.; Choi, Y.; and Kim, G. 2020. CurlingNet: Compositional Learning between Images and Text for Fashion IQ Data. *arXiv* .

Zhao, B.; Feng, J.; Wu, X.; and Yan, S. 2017. Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search. In *CVPR*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *ICCV*.