Fashion Image Retrieval with Text Feedback by Additive Attention Compositional Learning

Yuxin Tian[1], Shawn Newsam[1], Kofi Boakye[2]
[1]University of California, Merced, [2]Pinterest
{ytian8, snewsam}@ucmerced.edu, kofi@pinterest.com

Abstract

Effective fashion image retrieval with text feedback stands to impact a range of real-world applications, such as e-commerce. Given a source image and text feedback that describes the desired modifications to that image, the goal is to retrieve the target images that resemble the source yet satisfy the given modifications by composing a multi-modal (image-text) query. We propose a novel solution to this problem, Additive Attention Compositional Learning (AACL), that uses a multi-modal transformer-based architecture and effectively models the image-text contexts. Specifically, we propose a novel image-text composition module based on additive attention that can be seamlessly plugged into deep neural networks. We also introduce a new challenging benchmark derived from the Shopping100k dataset. AACL is evaluated on three large-scale datasets (FashionIQ, Fashion200k, and Shopping100k), each with strong baselines. Extensive experiments show that AACL achieves new state-of-the-art results on all three datasets.

1. Introduction

Image retrieval is a fundamental task in computer vision and serves as the cornerstone for a wide range of applications such as fashion retrieval [41, 53], geolocalization [40, 58], and face recognition [56]. There are several ways to formulate the search query such as keywords [2, 69], a query image [64, 62], or even a sketch [21, 34, 67, 8, 9, 51]. However, a core challenge in traditional image retrieval is that it is difficult for the user to refine the retrieved items based on their intentions. A range of approaches to incorporate user feedback to refine the retrieved images have been explored. Combining natural language feedback with a query image is a particularly promising framework since it provides a natural and flexible way for users to convey the image modifications that they have in mind.

In this work, we investigate image retrieval with text feedback where the goal is to retrieve images that are similar to a query image but incorporate the modifications described by the text. Such multi-modal and complementary input provides users with a powerful and intuitive visual search experience. However, as a multi-modal learning problem, it requires the synergistic understanding of both visual and linguistic content which can be a challenge. While image search with text feedback lies at the intersection of vision and language analysis, it differs from other extensively studied vision-and-language tasks, such as image-text matching [38, 36, 70, 28], image captioning [50, 47, 16], and visual question answering [22, 30, 12, 10]. This difference stems from the significant challenge of learning a *composite representation* that jointly captures the *visual* content of the query image and the *linguistic* information in the accompanying text to match the target image of interest.

Figure 1: We consider the task of retrieving new images that resemble the reference image while changing certain aspects as specified by text. Best viewed in color.

A fundamental challenge in image-text compositonal learning is characterizing global concepts from the query image and text representation simultaneously. For instance, when the text describes a modification to the color and neckline of a dress in a query image, the composition module should capture the concept of transforming the color and neckline, but it should also preserve the other visual concepts such as the trim, and material of the dress (Figure 1).

Another challenge is how to *selectively* modify the query image representation using the captured contextual information so that it is close to the target image representation in the latent space.

We propose a novel transformer-based Additive Attention Compositional Learning (AACL) model to address these challenges. The key idea is that we learn a contextual vector from the joint visiolinguistic representation. AACL then selectively modifies the query image tokens using the global context vector such that the composite features preserve the visual content of the image that should not be changed while transforming the relevant content according to the accompanying text.

We empirically compare our AACL approach with the state-of-the-art (SOTA) methods for visual search with text feedback on three large-scale fashion datasets: FashionIQ [23], Fashion200k [24], and a new challenging benchmark derived from Shopping100k [3]. We show that our proposed compositional learning method outperforms existing methods on all three datasets.

We make the following fundamental contributions:

- We propose a novel multi-modal additive attention layer capable of learning a global context vector which is used to selectively modify the image representation in an efficient way.
- We develop a fully transformer-based model for the challenging task of visual search with text feedback and demonstrate that it achieves state-of-the-art performance through extensive experiments on several large-scale fashion datasets.
- We create a new image-text retrieval dataset derived from Shopping100k. This new dataset features a wider range of fashion categories and attributes, resulting in an additional challenging benchmark for the research community.

2. Related Work

2.1. Image Retrieval with Text Feedback

Image retrieval with text feedback has been of interest to the computer vision research community for some time and a number of efforts (e.g., [5, 45, 60, 7]) have investigated effective ways to combine image and text representations. The text feedback can be provided in various ways, including absolute attributes (e.g., "red") [2, 69, 24], simple relative attributes (e.g., "more red") [48, 35, 65], or full natural language phrases [60, 4, 29, 14, 20, 55, 31]. Natural language is the preferred method of interaction between humans and computers in contemporary search engines. For image search in particular, it allows a user to convey detailed and precise specifications or modifications in a very natural way. We therefore focus on query-based image search with accompanying natural language phrases.

Previous methods [4, 13, 31, 20, 55] for image retrieval with text feedback rely heavily on convolution to aggregate features. In contrast, ours is the first approach to efficiently learn features globally via attention. Previous works have also relied on complicated hierarchical feature aggregation [14, 29], multiple forms of text feedback [14, 4], or multiple loss functions [14, 29, 4]. The winning solutions [31, 32, 54] for the FashionIQ 2020 challenge—an interactive image retrieval challenge—employed common performance boosting techniques such as careful hyperparameter tuning and model ensembles to improve the results. In contrast, AACL focuses on the *design of the image-text composition module* and achieves state-of-the-art performance via feature fusion in one step, which is more efficient and easier to adapt to other frameworks.

2.2. Image-Text Composition

While there has been much effort and different kinds of methods proposed to achieve the top scores on benchmarks involving images and text, relatively few have focused on the image-text composition module itself. In [33], the authors propose a multi-modal residual network (MRN) that learns representations by fusing visual and textual features through element-wise multiplication and residual learning. FiLM [49] utilizes a linear modulation component in which text information modifies the image representation via a feature-wise affine transformation. Vo *et al.* proposed TIRG [60], which uses a gating mechanism to determine the channels of the image representation that should be modified by the conditioning text. In ComposeAE [4], a complex embedding space that semantically ties the representations from text and image modalities is designed. Recently, MAAF [20] improved multi-modal image search via a Modality-Agnostic Attention Fusion model. This model uses a dot product attention mechanism as found in the standard transformer architecture. Additionally, resolution-wise pooling is proposed to aggregate fine-grained features from a ResNet [25] CNN. RTIC [55] consists of a residual text and image composer to encode the errors between the source and target images in the latent space and includes a graph convolutional network for regularization. Our work differs from these composition modules in that we utilize a novel image and text composition module via additive attention [6, 46] to model global contexts. Furthermore, we use an element-wise product to model the interaction between the global context and each input token, which both greatly reduces the computational cost and effectively captures the contextual information [33, 31, 63].

2.3. Attention Mechanism

The concept of attention has gained popularity recently in neural networks as it allows the models to learn representations from different modalities [33, 27, 20, 14, 5, 18]. The two most commonly used attention functions are additive [6], and dot-product (multiplicative) attention [59].

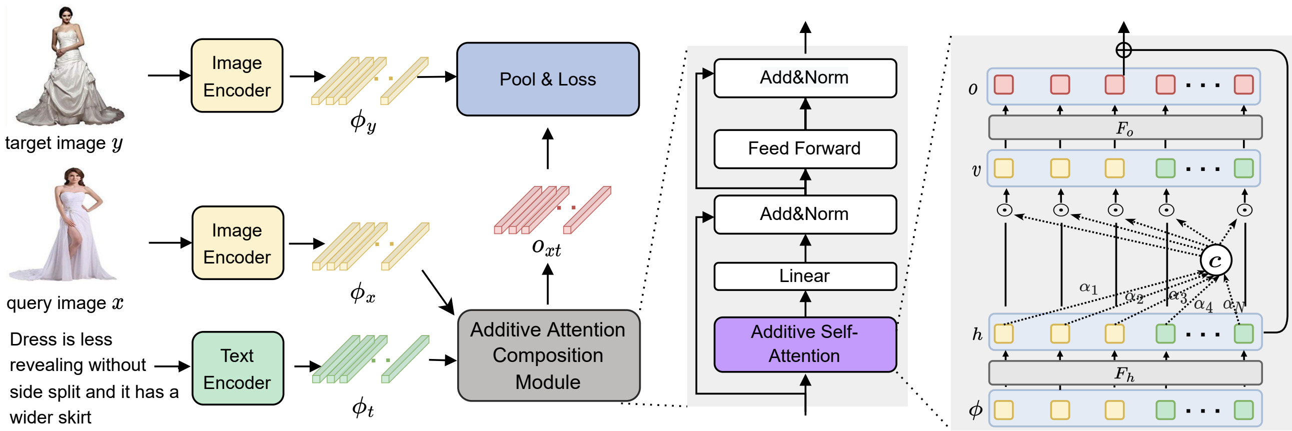


Figure 2: Overview of our Additive Attention Compositional Learning framework. Given a pair of query image and text as input, our goal is to learn a composite representation that aligns to the target image representation. AACL contains three major components: an image encoder (Sec. 3.1), a text encoder (Sec. 3.1), and an Additive Attention Composition Module (Sec 3.2) that can be plugged into different models for feature fusion. "$\odot$" represents Hadamard product.

Dot-product attention has a drawback, however, in that it has to attend to all the tokens on the source side for each target token, which is expensive and can potentially be impractical for longer sequences. Additive attention has been shown experimentally to achieve higher accuracy than multiplicative attention in some scenarios [46, 63]. Inspired by this, we propose an *additive attention composition module* for feature fusion.

2.4. Vision-Language (VL) Pre-training

Although image retrieval with text feedback shares some similarity with VL pre-training [57, 15, 39, 68, 66, 37], the focus of our work is distinct. The goal of VL pre-training is to learn cross-modal representations, which can be adapted to serve various down-stream tasks via fine-tuning [39]. However, our work focuses on the image-text composition module itself, which performs single stage late feature fusion with image and text embeddings from separate transformer encoders.

3. Method

Figure 2 presents the overall architecture of our Additive Attention Compositional Learning (AACL) framework. Given a source image x and text feedback t as the input query, the goal of AACL is to learn a composite representation o_{xt} that can be used to retrieve relevant images y from a target database. AACL contains three key components: (1) an image encoder for visual semantic representation learning, (2) a text encoder for natural language representation learning, and (3) an additive attention composition module that modifies the source image representation according to the text representation. In contrast to other approaches that use multiple stages of feature composition and matching (e.g., [14]), AACL does this in one stage using the final output of the image and text encoders.

In the following, we first provide an overview of the two encoders in Section 3.1. We then detail our novel composition module in Section 3.2 and our model optimization in Section 3.3.

3.1. Image and text representation

Image Representation: We employ a Swin Transformer [44] to derive a discriminative representation of the visual content of an image. As a transformer inherently learns visual concepts of increasing abstraction in a compositional, hierarchical order, we conjecture that image features from the final layer may not fully capture the visual information of the lower levels. We thus concatenate image tokens extracted from the final (Stage 4) and penultimate (Stage 3) layers of the Swin Transformer. Unless otherwise specified, our model uses these $49 + 49 = 98$ image tokens for multi-level image understanding. A learned linear projection maps each image token to d dimensions so that the final image representation is $\phi_x \in \mathbb{R}^{98\times d}$.

Text Representation: The DistilBERT language representation model [52] is used to encode the semantics of the accompanying text. DistilBERT naturally yields m tokens for the input words, namely the hidden states of the last layer of the model. We concatenate these tokens to form the final text representation $\phi_t \in \mathbb{R}^{m\times d}$.

3.2. Additive Attention Composition Module

In order to jointly represent the image and text components of the query, we seek to transform the visual features conditioned on language semantics. To accomplish this, we propose an additive attention composition module for feature fusion. This module consists of multiple composition blocks that each employ additive self-attention to learn a context vector which then selectively modifies the joint visiolinguistic representation. The final output of these blocks yields a modified image representation that is meant to faithfully capture the input image and text information.

Visiolinguistic Representation: In order to obtain the input representation for our first composition block, the image tokens ϕ_x and text tokens ϕ_t are concatenated to obtain the visiolinguistic representation $\phi_{xt} = [\phi_x, \phi_t]$. The final representation is denoted as $\phi_{xt} \in \mathbb{R}^{N \times d}$, where N is the combined count of image and text tokens.

Additive Self-Attention Layer: In order to discover the latent relationships essential for learning the transformation, we use the additive attention mechanism to learn a context vector c, then selectively suppress and highlight the representations from each token. Similar to [63], we first use a linear transformation layer to transform the input sequence into the hidden states: $h = \mathcal{F}_h(\phi_i), i \in N$. The context vector c that is learned to modify each token is generated as a weighted sum of these tokens h_i:

$$c = \sum_{i=1}^{N} \alpha_i h_i. \tag{1}$$

The weight α_i of each token h_i is computed by

$$\alpha_i = \frac{\exp\left(\mathbf{w}_h^T \mathbf{h}_i / \sqrt{d}\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{w}_h^T \mathbf{h}_j / \sqrt{d}\right)}, \tag{2}$$

where $\mathbf{w}_h \in \mathbb{R}^d$ is learned during the training process, and $\mathbf{w}_h^T \mathbf{h}_j$ scores how much each input token contributes to the global context.

Next, to selectively suppress and highlight the visual content in h, a Hadamard product is introduced to reuse the global contextual information, which is motivated by its effectiveness in modeling the nonlinear relationship between two vectors [61, 63, 26]. It is formulated as $v_i = c \odot h_i$. Another linear transformation layer $\mathcal{F}_o$ is applied to each token v_i to learn its hidden representation. To form the final output of the additive attention layer, we add the hidden states h_i that capture relevant source-side information to the transformed latent features. The final output of the additive self attention layer is:

$$o_i = h_i + \mathcal{F}_o\left(c \odot h_i\right) \tag{3}$$

Composition Block: Following the standard transformer architecture [59], the additive attention composition module is composed of a stack of L identical blocks with multiple heads. Different attention heads use the same formulation but different parameters, which allows the model to jointly attend to information from different representation subspaces at different positions. Each block has an additive self-attention layer followed by a linear layer and a feed-forward neural network. We also employ a residual connection and layer normalization after these linear and feed-forward components to get the composited image-text representation o_{xt}.

3.3. Deep Metric Learning

Our objective during training is to push the "modified" image representation o_{xt} and the target image representations ϕ_y closer, while pulling apart the representations of dissimilar images. A batch-based classification loss as in [60] is used to train the model as early experiments showed that the triplet loss performs worse for the Recall@k metric. Each batch is constructed from N pairs of a query (image and text) and its corresponding target image.

$$L = \frac{1}{B} \sum_{i=1}^{B} -\log \left\{ \frac{\exp\left\{\kappa\left(\phi_y, o_{xt}\right)\right\}}{\sum_{j=1}^{B} \exp\left\{\kappa\left(\phi_y, o_{xt}\right)\right\}} \right\} \tag{4}$$

where B is the batch size and κ is a similarity kernel that is implemented as the dot product in our experiments.

4. Experiments

4.1. Experimental Setup

Datasets: We evaluate our model on three datasets—FashionIQ, Fashion200k and our modified version of Shopping100k—in order to validate its ability to generalize to a variety of natural language expressions. We provide details of these datasets in Sections 4.2, 4.3, and 4.4, respectively.

Implementation Details: We use the PyTorch deep learning framework to conduct all our experiments. The Swin Transformer [44] is used as the backbone for the image encoder. The transformer model is initialized using weights first pre-trained on ImageNet-22K and then fine-tuned on ImageNet-1K [17].

We extract sequences of 1024-dimensional tokens from Stages 3 and 4 of the model and then project the tokens to d dimensions, which for our experiments is 768. We learn the text embedding using a pre-trained DistilBERT model [52], which yields a 768-dimensional token for each input word. The original BERT model is pre-trained on BooksCorpus (800M words) and English Wikipedia (2,500M words) [19]. We employ 3 additive attention composition blocks and 8 parallel attention heads for each block. For training, we use SGD optimization with a learning rate of 0.035. We train all models using 4 GPUs with a batch size of 32 per GPU. For FashionIQ, we employ a learning rate decay of 0.1 every 10 epochs for 60 epochs. For Fashion200k and our modified Shopping100k, we use the same decay value but every 30 epochs with a total of 100 epochs. We report the average and standard deviation of five trials for all our experiments to obtain more meaningful results.

Evaluation Metric: Following [60, 55, 20], we adopt Recall@K (denoted as R@K for short) for evaluation, a standard metric in retrieval. The relative margin expresses the absolute changes as a percentage of the baseline value.

Compared Methods: We compare the results of AACL with several methods, namely: FiLM, MRN, TIRG, Com-

Table 1: Comparison of image search with text feedback on FashionIQ. Averaged R@10/50 computed over all three categories. * denotes results obtained with the same image encoder and text encoder as AACL.

Model	Shirt		Dress		Toptee		Average
	R@10	R@50	R@10	R@50	R@10	R@50	(R@10 + R@50)/2
MRN [33]	15.88	34.33	12.32	32.18	18.11	36.33	24.86
FiLM [49]	15.04	34.09	14.23	33.34	17.30	37.68	25.28
TIRG [60]	16.12	37.69	19.15	43.01	21.21	47.08	30.71
ComposeAE [4]	9.96	25.14	10.77	28.29	12.74	30.79	19.61
MAAF [20]	21.30	44.20	23.80	48.60	27.90	53.60	36.57
RTIC [55]	22.03	45.29	27.37	52.95	27.33	53.60	38.10
TIRG*	21.38±0.54	46.28±0.78	25.82±0.39	53.21±0.33	26.73±0.72	53.17±0.29	37.77±0.21
MAAF*	23.55±0.31	46.38±1.34	28.75±0.63	54.48±0.49	29.70±0.45	55.84±0.87	39.78±0.68
RTIC*	23.03±0.63	46.68±0.52	26.86±0.74	52.80±0.61	27.21±0.89	53.24±0.66	38.31±0.67
AACL	**24.82±0.62**	**48.85±0.77**	**29.89±0.65**	**55.85±0.87**	**30.88±1.2**	**56.85±1.16**	**41.19±0.88**

Figure 3: Qualitative results of AACL on FashionIQ dataset. Blue/green boxes: query/target images.

poseAE, MAAF and RTIC. We explained them briefly in Section 2.2.

4.2. FashionIQ

FashionIQ [23] is a natural language based interactive fashion product retrieval dataset. It contains 77,684 images crawled from Amazon.com, covering three categories: Dresses, Tops&Tees and Shirts. Among the 46,609 training images, there are 18,000 image pairs. Each pair is accompanied by on average two natural language sentences that describe one or multiple visual properties to modify in the reference image, such as *"is shiny"* or *"is blue in color and floral, and with white base"*. We follow the same evaluation protocol as [23], using the same training split and evaluating on the validation set. We report results on individual categories, as well as the average results over all three.

Table 1 compares the performance of AACL and the other methods on FashionIQ. We observe that AACL is superior to all reported results by a large margin (top half). AACL even outperforms methods that include factors other than the composition module itself, such as the target image captions, model ensembles, and additional joint loss functions [4]. We further note that AACL is actually complementary to some of these methods and could, in fact, be used as their composition modules. For a like-to-like fair comparison, we also reproduced the best competitors, focusing on just the composition module itself. That is, we utilized the same image and text encoders—namely, Swin Transformer and DistilBERT—and the same optimizer. In this scenario AACL surpasses TIRG, RTIC, and MAAF by an overall margin of 3.42%, 2.88% and 1.41% respectively in average R@10 and R@50 scores. Figure 3 presents our qualitative results on FashionIQ. We show top-5 retrieved images for each query image-text pair. These results demonstrate that our model can handle complex and realistic text descriptions. We also observe that our model can jointly comprehend global appearance (e.g., colors, material), as well as local fine-grained details (e.g., straps and neckline, length of sleeves), for image search.

4.3. Fashion200k

Fashion200k [24] is a large-scale fashion dataset crawled from multiple online shopping websites. It contains more than 200k fashion images collected for attribute-based product retrieval covering five categories, namely, Dresses, Jackets, Pants, Skirts, Tops. It also covers a diverse range of fashion concepts, with a total vocabulary size of 5,590. Each image is tagged with descriptive text corresponding to a product description, such as *"beige v-neck bell-sleeve top"*. Following [60], we use the training split of 172,049 images for training and the test set of 33,480 test queries for evaluation. During training, pairwise images with attribute-like modification texts are generated by comparing their product descriptions on-the-fly, e.g., *"replace black with blue"* or *"replace mini with midi"*.

Table 2 shows our model achieves compelling results compared to other methods, most notably for R@1 where AACL outperforms the best competitor MAAF by a relative margin of 9.4%. We also observe that token based methods, namely MAAF and AACL, perform better than residual based methods. This indicates that the rich information contained in tokens is beneficial for feature composition. Figure 4 shows our qualitative results on Fashion200k. Our model is able to retrieve new images that resemble the reference image, while changing certain attributes conditioned

Table 2: Comparison of image search with text feedback on Fashion200k dataset. * denotes our implementation results obtained with the same image encoder and text encoder as AACL.

Model	R@1	R@10	R@50
FiLM [49]	12.9	39.5	61.9
MRN [33]	13.4	40.0	61.9
TIRG [60]	14.1	42.5	63.8
ComposeAE [4]	16.5	45.4	63.1
DCNet [31]	–	46.9	67.6
MAAF [20]	18.94	–	–
TIRG*	17.22±0.39	56.52±1.85	75.60±0.09
MAAF*	17.79±0.98	57.57±0.98	77.51±0.63
RTIC*	17.05±0.96	54.65±0.79	75.54±1.63
AACL	**19.64±1.66**	**58.85±1.01**	**78.86±0.43**

Figure 4: Qualitative results of AACL on Fashion200k dataset. Blue/green boxes: query/target images.

Table 3: Number of images in select categories (count > 2k) in Shopping100k dataset.

Jacket	Shirt	T-shirt	Jumper	Shorts	Trouser	Jean	Swim	Bottoms[1]	Skirt	Dress
7,528	14,853	22,071	11,797	5,099	4,630	6,229	5,497	3,726	2,528	12,119

on text feedback—e.g., fit, color and length. We also observe that all retrieved images share the same semantics and are visually similar to the target image, indicating the quantitative performance is potentially underestimated.

4.4. Shopping100k

Shopping100k [3] is a large-scale fashion dataset of individual clothing items extracted from different e-commence providers. It contains 101,021 images of 12 fashion attributes, covering the following categories: "collar", "color", "fabric", "fastening", "fit", "gender", "length", "neckline", "pattern", "pocket", "sleeve length", and "sport". A total of 151 different labels are generated by combinations of different attributes and the corresponding attributes values. Compared to FashionIQ and Fashion200k, the Shopping100k dataset is more diverse and only contains garments in isolation. In addition, FashionIQ and Fashion200k only contain 3 and 5 apparel categories, respectively.

Each image in Shopping100k is tagged with the attributes and attribute values, such as *"Neckline: Backless, Sleeve: 3/4, Color: Navy, Fabric: Jersey, Pattern: Print, Category: Shirt, Fit: Large, Gender: Female"*. There are 15 high-level apparel categories. To generate the dataset for image retrieval with text feedback, we remove categories that contain fewer than 2,000 images, namely "coat", "suit", "jumpsuit", "pyjamas", and "tracksuit". The final set of 11 categories is listed in Table 3 along with the number of images in each category. A training split with 76,867 images and a validation split with 19,210 images is randomly sampled from these remaining categories.

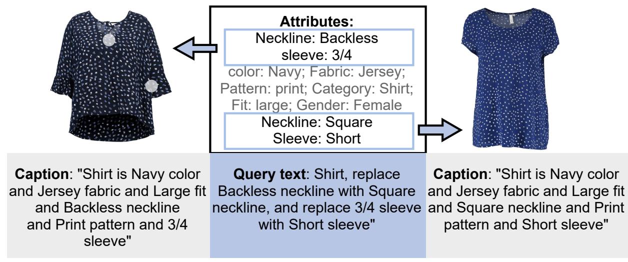


Figure 5: Example of image pair and generated text query from Shopping100k dataset. Gray words indicate shared attributes.

To generate the training image pairs and modification text, we first derive a descriptive caption for each image using its tagged attribute values by concatenating the category with "is", followed by attributes joined by "and"—e.g., *"Shirt is Navy color and Jersey fabric and Large fit and Backless neckline and Print pattern and 3/4 sleeve"*. Queries are created by selecting image pairs that differ in two attributes in the description. Note that we constrain the image pairs to be from the same apparel category and gender. The modification text is created with the apparel category plus the attribute modifications following the pattern "replace xx with xx"—i.e. *"Shirt, replace Backless neckline with Square neckline, and replace 3/4 sleeve with Short sleeve."* (Figure 5). During training, the query and target image pairs are selected on-the-fly based on the number of attributes we specify. For our experiments, 16,237 fixed test query pairs are generated from the validation set for performance evaluation.

Table 4 compares our approach to other methods on Shopping100k. Our model is shown to clearly outperform the SOTA baselines. Figure 6 presents some qualitative examples. These examples yield three observations. First, our model is capable of understanding rich image-text representations, including global attributes such as color, pattern, and fit, as well as local attributes such as collar, neckline, and sleeves. Second, our model is capable of using the text information to selectively modify the query images. As an example, for the first query the retrieved images preserve the striped pattern even though it is not requested in

[1]Full name of category "Bottoms" is "Tracksuit Bottoms".

Table 4: Comparison of image search with text feedback on our modified Shopping100k dataset. Averages are computed over all categories. * denotes our implementation results obtained with the same image encoder and text encoder as AACL.

Model	Dress	Jacket	Jean	Jumper	Shirt	Shorts	Skirt	Swimming	T-shirt	Bottoms	Trouser	Average
Recall@1												
TIRG*	6.81±0.58	10.46±0.97	4.83±1.43	11.87±1.26	13.15±1.25	12.38±1.16	10.92±1.22	13.51±1.49	11.87±0.80	8.32±0.60	13.03±1.77	10.65±0.37
MAAF*	7.05±0.86	12.43±0.76	5.79±1.34	13.19±0.88	**14.44±1.28**	13.21±1.68	12.11±0.77	12.41±0.71	12.89±1.16	**10.28±1.35**	12.89±0.87	11.52±0.39
RTIC*	6.80±0.09	11.70±0.90	5.27±0.90	12.08±1.39	13.93±1.33	11.83±0.97	10.96±1.44	13.18±0.99	12.60±0.99	8.49±0.65	11.70±1.70	10.78±0.44
AACL	**7.70±0.67**	**12.63±0.93**	**7.27±0.96**	**13.30±0.31**	14.21±0.52	**14.38±1.14**	**14.55±1.22**	**16.22±1.02**	**13.66±0.28**	10.00±0.53	**14.14±0.63**	**12.55±0.32**
Recall@10												
TIRG*	34.22±0.53	49.86±0.47	29.23±0.48	51.08±0.89	50.22±0.72	50.43±0.52	55.85±0.58	51.86±1.49	47.19±1.04	41.69±0.59	51.06±1.28	46.61±0.35
MAAF*	35.01±1.85	51.48±1.67	**31.78±1.12**	51.70±2.45	52.15±1.96	50.64±1.30	54.70±3.36	54.74±2.46	**49.31±1.79**	44.00±2.87	52.08±0.63	47.96±0.65
RTIC*	33.17±1.92	50.51±2.11	29.21±4.36	48.92±3.39	50.90±2.89	50.29±0.74	51.96±2.09	51.62±2.02	46.71±2.41	42.24±1.31	51.46±1.25	46.09±1.03
AACL	**35.16±0.54**	**51.63±1.33**	30.80±1.79	**52.31±0.89**	**52.52±1.32**	**54.63±1.66**	**57.54±0.95**	**56.13±2.13**	49.18±1.40	**46.69±1.06**	**54.63±1.72**	**49.20±0.46**
Recall@50												
TIRG*	66.15± 0.80	81.50±0.38	62.47±0.19	80.74±2.40	82.43±0.28	81.36±0.95	85.57±1.66	83.91±1.20	79.32±1.81	77.94±1.18	85.02±1.35	78.76± 0.69
MAAF*	68.42±1.42	82.73±2.29	63.24±2.94	82.28±1.36	84.41±1.90	82.06±1.66	88.19±0.78	**85.32±2.27**	**81.07±1.34**	81.17±0.67	86.75±0.82	80.51±0.56
RTIC*	67.30±2.12	81.92± 2.42	**64.30±5.31**	80.27±2.37	83.45±1.58	82.22±1.88	84.71±1.57	84.15±2.46	78.87±1.95	79.47±0.88	85.37±1.92	79.27±1.12
AACL	**69.21±0.37**	**83.30±1.77**	63.92±3.59	**82.30±0.36**	**84.75±1.21**	**85.50±1.30**	**88.94±0.78**	85.31±1.52	80.54±1.18	**82.83±0.88**	**87.61±0.76**	**81.29±1.11**

Figure 6: Qualitative results of AACL on Shopping100k dataset. Blue/green boxes: query/target images.

Table 5: Ablation of using tokens from different Swin Transformer stages on our modified Shopping100k dataset.

Stage(s)	Recall@1	Recall@10	Recall@50
Stage 2 + 3 + 4	11.92	48.78	80.74
Stage 3 + 4	**12.26**	**49.20**	**81.29**
Stage 4	12.01	48.56	81.25

the text feedback. Five of the top-5 retrieved candidates fulfill the "long sleeves" requirement and four candidates have "low-v-neck". Third, the model is capable of capturing minor modifications such as "kent collar" *vs.* "mandarin collar", suggesting it can be successfully utilized in fine-grained search.

4.5. Ablation Study

Image representation: Table 5 compares the performance of AACL when using different image representations from the Swin Transformer on our modified Shopping100k dataset. The experiments reveal that using image tokens from Stages 3 and 4 is most effective for this task. The concatenation of two stages from the encoder considers richer forms of image representation. Somewhat surprisingly, concatenating representations from Stage 2 does not seem to benefit the task. This may suggest that at some point, the lower level information may distract the model from capturing meaningful global contextual information.

Additive attention: To assess the importance of additive attention, we perform a comparison by substituting with dot-product attention. Table 6 "Additive→Dot-Product" shows the comparison on our modified Shopping100k dataset. From these results, we that AACL does benefit consistently from the additive attention. In addition, dot product attention is more computationally expensive than additive attention ($O(n^2)$ *vs.* $O(n)$) and as such the benefits of additive attention extend beyond evaluation performance gains.

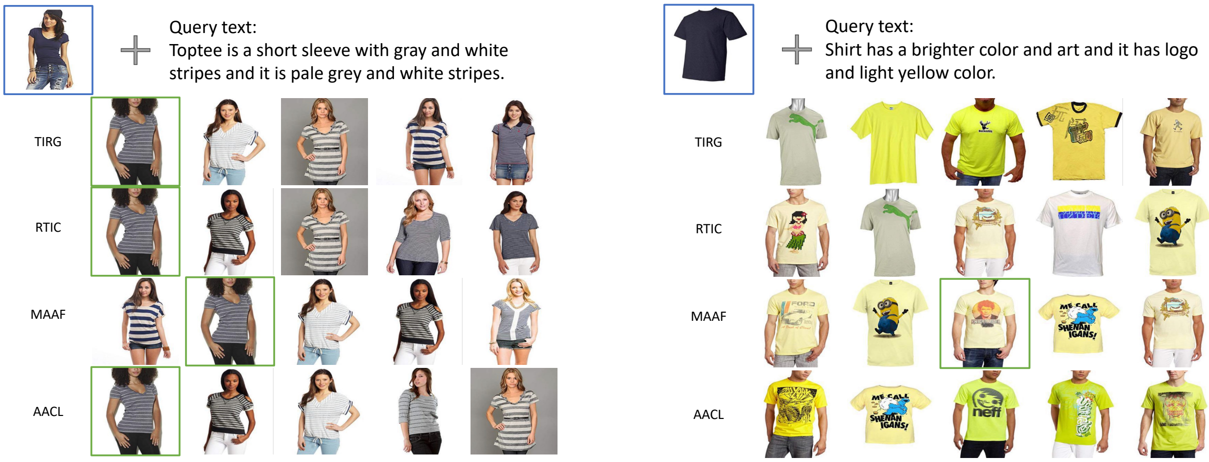
Table 6: Ablation of self-attention layer on our modified Shopping100k dataset. We separately examine substituting additive self-attention with standard dot-product and changing the Hadamard product to addition.

Method	Recall@10	Recall@50
Additive→Dot-Product	48.37	80.14
Product→Addition	48.56	80.45
AACL	**49.20**	**81.29**

Interaction function: We study the effect of using different functions, namely addition and Hadamard product, to model the interactions between the context vector and the individual tokens. We compare the standard AACL and this variant on Shopping100k. The results are shown in Table 6 "Product→Addition". The Hadamard product performs consistently better than addition, indicating this form of non-linear modeling is beneficial.

4.6. Additional Qualitative Results

Figure 7 qualitatively compares our AACL model with TIRG, RTIC and MAAF on the FashionIQ dataset. Note that the query text of FashionIQ most closely resembles natural language as the queries are provided by annotators from English-speaking countries. Even though for each query image a single target image is defined, there can be multiple "perceptually acceptable" images. This is because there may exist multiple items in the database that are similar to the target image and satisfy the modifying text component of the query. In Figure 7a, for example, there is more than one toptee that is short sleeved with gray and white stripes among the retrieved items, but only the target image is considered a correct match. Compared to the other models considered, our AACL model tends to find the best matching images that satisfy all conditions in the queries. In



(a) Successful examples (b) Failed examples

Figure 7: Qualitative comparison on FashionIQ dataset. We present the query image and query text in the first row, followed by the top-5 retrieved images from the various models in subsequent rows. Blue/green boxes: query/target images.

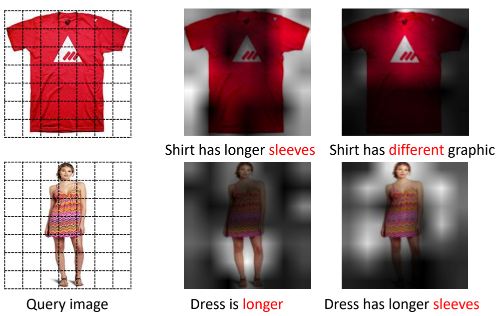


Figure 8: Attention visualization of AACL model on FashionIQ dataset. Words with highest attention value in red.

contrast, Figure 7b shows a failure case. Here, our AACL retrieves several "perceptually acceptable" results, though, this is treated as a failed case.

To interpret the attention learned by AACL, we visualize the attended regions in Figure 8. We apply a mask based on the attention flow to the input query image. The attention flow is generated as follows: We first multiply the α_i in Equation 2 across all blocks to get the total attention flow for each token. Subsequently, the minimum word token flow score is mapped to zero and the maximum to one. Note that, since we are using the Swin Transformer as the image encoder, the encoded feature maps are 7×7 and the resulting visualization resolution appears lower than with other models. Nevertheless, given the same query image, we do observe that the spatially attended regions vary with different query text. This indicates that the additive self-attention selects different visual content to transform conditioned on the text query.

4.7. Limitations

The retrieved images are, to some extent, limited by what images are present in the target datasets. We note that the retrieved images may not fully fulfill the desired changes described by the text modifier while keeping the rest of the query image the same if there is no such target images. Another limitation is the attention visualization. As an active research topic, current attention visualization methods mainly focus on dot-product attention [1, 11]. Those widely adopted methods are not compatible with our additive attention module, and as such we adopted a simpler—and potentially less precise—visualization approach. How to obtain accurate labeled data is critical for the success of training models [42, 43]. However, the template-based relative caption generation method, although widely used, is not as accurate and diverse as human annotations.

5. Conclusion And Future Work

We present AACL, a novel and general-purpose solution to the challenging task of image search with text feedback. This framework features an additive self-attention layer that selectively preserves and transforms multi-level visual features conditioned on text semantics to derive an expressive composite representation. We validate the efficacy of AACL on three datasets, and demonstrate its consistent superiority in handling various text feedback for natural language expression. Overall, our work provides a novel approach along with a comprehensive evaluation, which collectively advance the research in interactive visual search using text feedback.

In addition to addressing some limitations mentioned above, there are many possible future research directions. First of all, we plan to leverage the recent advances in image generation to create realistic and desired images based on the query image-text pair. Second, automated relative captioning can be applied to generate text modifiers that better resemble natural language and reduce noisy query text.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[2] Kenan E. Ak, Ashraf A. Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Kenan E. Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A. Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[4] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[5] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *arXiv*, 2014.

[7] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[8] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] Ayan Kumar Bhunia, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Remi Cadene, Hedi Ben-Younes, Nicolas Thome, and Matthieu Cord. Murel: Multimodal Relational Reasoning for Visual Question Answering. In *CVPR*, 2019.

[11] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[12] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Proceedings of the european conference on computer vision (ECCV)*, 2020.

[14] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[16] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[18] Xueqing Deng, Yi Zhu, Yuxin Tian, and Shawn Newsam. Scale aware adaptation for land-cover classification in remote sensing imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[20] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020.

[21] Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven J. Rennie, and Rogério Schmidt Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019.

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[28] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: Multimodal, multitask retrieval across languages, 2021.

[29] Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy. Sac: Semantic attention composition for text-conditioned image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[30] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual Compositional Learning in Interactive Image Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence)*, 2021.

[32] Jongseok Kim, Youngjae Yu, Seunghwan Lee, and GunheeKim. Cycled compositional learning between images and text. In *arXiv*, 2021.

[33] 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 *Advances in Neural Information Processing Systems*, 2016.

[34] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[35] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 2015.

[36] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[38] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[39] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. 2020.

[40] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] Yen-Liang Lin, Son Tran, and Larry S. Davis. Fashion outfit complementary item retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[42] Jie Liu, Jiawen Liu, Zhen Xie, Xia Ning, and Dong Li. Flame: A self-adaptive auto-labeling system for heterogeneous mobile processors. In *2021 IEEE/ACM Symposium on Edge Computing (SEC)*, 2021.

[43] Miaomiao Liu, Xianzhong Ding, and Wan Du. Continuous, real-time object detection on mobile devices without offloading. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, 2020.

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019.

[46] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[47] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[48] Devi Parikh and Kristen Grauman. Relative attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2011.

[49] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence)*, 2018.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[51] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[52] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[53] Raymond Shiau, Hao-Yu Wu, Eric Kim, Yue Li Du, Anqi Guo, Zhiyuan Zhang, Eileen Li, Kunlong Gu, Charles Rosenberg, and Andrew Zhai. Shop the look. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[54] Minchul Shin, Yoonjae Cho, and Seongwuk Hong. Fashion-iq 2020 challenge 2nd place team's solution, 2020.

[55] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. Rtic: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*, 2021.

[56] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[57] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[58] Yuxin Tian, Xueqing Deng, Yi Zhu, and Shawn Newsam. Cross-time and orientation-invariant overhead image geolocalization using deep local features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[60] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[61] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. Item silk road. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.

[62] Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian triplet loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[63] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Fastformer: Additive attention can be all you need. *arXiv preprint arXiv:2108.09084*, 2021.

[64] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[65] Aron Yu and Kristen Grauman. Thinking outside the pool: Active training image creation for relative attributes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[66] Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L. Berg, and Ning Zhang. Commercemm. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[67] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[68] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models, 2021.

[69] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[70] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.