Image Search with Text Feedback by Visiolinguistic Attention Learning

Yanbei Chen*
Queen Mary University of London
yanbei.chen@qmul.ac.uk

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Loris Bazzani
Amazon
bazzanil@amazon.com

Abstract

Image search with text feedback has promising impacts in various real-world applications, such as e-commerce and internet search. Given a reference image and text feedback from user, the goal is to retrieve images that not only resemble the input image, but also change certain aspects in accordance with the given text. This is a challenging task as it requires the synergistic understanding of both image and text. In this work, we tackle this task by a novel Visiolinguistic Attention Learning (VAL) framework. Specifically, we propose a composite transformer that can be seamlessly plugged in a CNN to selectively preserve and transform the visual features conditioned on language semantics. By inserting multiple composite transformers at varying depths, VAL is incentive to encapsulate the multi-granular visiolinguistic information, thus yielding an expressive representation for effective image search. We conduct comprehensive evaluation on three datasets: Fashion200k, Shoes and FashionIQ. Extensive experiments show our model exceeds existing approaches on all datasets, demonstrating consistent superiority in coping with various text feedbacks, including attribute-like and natural language descriptions.

1. Introduction

Image search is a fundamental task in computer vision. It has been serving as the cornerstone in a wide range of application domains, such as internet search [42], fashion retrieval [34], face recognition [57] and product identification [44]. The most prevalent paradigms in image search take either image or text as the input query to search for items of interest, commonly known as image-to-image [15] and text-to-image matching [12]. However, an intrinsic downside of these paradigms lies in the infeasibility to refine the retrieved items tailored to users' intentions, especially when users cannot precisely describe their intentions by a single image or with all the keywords.

To overcome the aforementioned limitation, different user interactive signals have been explored over the past two decades [61]. The basic idea is to incorporate user feedback to refine or discover image items retrieved by the system [52, 81, 70, 11, 45, 28, 27, 18, 79, 2, 39, 16, 48, 75, 17]. Most of these interactions are delivered in the form of *text*, describing certain attributes [18, 79, 2] or relative attributes [45, 28, 75] to refine or modify upon a reference image. More recently, natural language feedback [17] is introduced as a more flexible way to convey user's intention for interactive image search. Despite having great potential value in practice, incorporating various types of text feedback for image search still remains understudied.

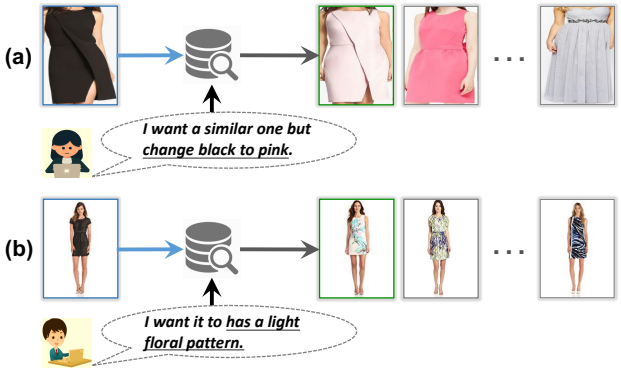


Figure 1. Given a *reference image* and *user text* as input, we consider the task of retrieving new images that resemble the reference image while changing certain aspects as specified by text. The text generally describes the visual content to refine in reference image, such as **(a)** a concrete attribute or **(b)** more abstract properties.

In this work, we investigate the task of image search with text feedback, which entitles user to interact with the system by selecting a reference image and providing additional text to refine or modify the retrieval results. Unlike the prior works that mostly focus on one type of text feedback, we consider the more general form of text, which can be either *attribute-like description*, or *natural language expression*. This poses a more challenging multimodal learning problem that requires the synergistic understanding of both visual and linguistic contents at different granularities – the given text may convey multi-granular semantics, ranging from *a concrete attribute* to *highly abstract visual properties* (Fig. 1). As a task lying at the intersection of vision and language, image search with text feedback, however, differs

*Work partially done during an internship with Amazon.

greatly from other extensively studied vision-and-language tasks, such as image-text matching [12, 73, 67], image captioning [24, 60], and visual question answering [4, 60]. This is because, it uniquely entails learning a composite representation that can jointly capture visual cues and linguistic information to match the target image of interest.

One intrinsic challenge is the difficulty to simultaneously *preserve* and *transform* the visual content in accordance with the given text. For instance, when a text snippet specifies the colour to modify (Fig. 1**(a)**), it means the other visual cues such as silhouette, pattern, trim should all be preserved in the retrieved items, with only the colour transformed to the desired one. Another challenge is to learn a composite representation that can jointly encapsulate visual and linguistic contents from *coarse* to *fine*-grain. Since the text feedback may convey multi-level semantics (Fig. 1), the composite representation is also expected to capture the multi-granular visiolinguistic information. To address these challenges in a unified solution, we propose a novel **V**isiolinguistic **A**ttention **L**earning (**VAL**) framework, which fuses vision and language features via attention learning at varying representation depths.

Briefly, VAL is featured with multiple composite transformers plugged at multi-level inside a CNN to compose visual features and language semantics. Our core idea is to learn the *attentional transformation* and *preservation* concurrently, such that the composite features not only *preserve* the unaltered visual content in image, but also *transform* certain content as specified by text. To train our VAL, we devise a *hierarchical matching* objective, which incentivises exclusive alignments to the desired visual and semantic features for discriminative feature learning.

To summarise, our **contribution** is two-fold:

- We tackle the challenging task of image search with text feedback by a novel Visiolinguistic Attention Learning (VAL) framework. VAL is characterised by multiple *composite transformers* that compose multi-level visual features and language semantics via *attention learning*. Through a hierarchical matching objective, VAL is incentive to encapsulate visual and linguistic contents as composite representations for effective image search.
- We set a new state-of-the-art on three datasets: Fashion200k, Shoes, and FashionIQ. Remarkably, VAL performs consistently well in coping with various types of text feedback, demonstrating a greater potential in practical use. We also present an insightful ablation study to analyse the underlying attentions learnt by VAL.

2. Related Work

Interactive image search aims to incorporate user feedback as an interactive signal to navigate the visual search. In general, the user interaction can be given in various formats, including relative attribute [45, 28, 75], attribute [79, 18, 2], attribute-like modification text [66], natural language [16, 17], spatial layout [37], and sketch [76, 74, 14]. As text is the most pervasive interaction between human and computer in contemporary search engines, it naturally serves to convey concrete information that elaborates user's intricate specification for image search. In this work, we investigate various text feedbacks for image search. Thanks to the rich annotations released recently on several fashion benchmark datasets [18, 17], we present the *first attempt* to consider richer forms of text feedback in *one-turn* interactive search, including *attribute-like* and *natural language* expression.

Attention mechanism is widely adopted as an important ingredient in various vision-and-language tasks, which aims to mimic human's capability of attending to salient sensory information [7]. To steer where to fixate in images, spatial attention is commonly used to assign importance weights on image regions. This helps to select informative regions for captioning [65, 3], or locate relevant visual content for question answering [72, 82]. For attention learning in vision and language domains, co-attention [36, 41] is generally adopted to fuse visual and textual contents by generating attention weights on image regions and question words. Recently, several self-attention mechanisms are proposed for VQA [77, 13, 23, 35], which builds upon transformer [64] to learn the inter-modal or intra-modal latent attention. Inspired by this line of works, we propose a generic visiolinguistic attention learning scheme, which learns the attentional interactions upon the visiolinguistic features. Unlike previous works that rely heavily on off-the-shelf Faster R-CNN [51] to extract image region features, our approach avoids the dependency on a pre-trained object detector, and thus generalises well to fine-grained visual search, especially when the imagery data does not share the common objects as those in the object detection datasets.

Composition learning is deemed as an essential functionality to build intelligent machine [29, 30]. The general aim is to learn a feature encoding that encompasses multiple primitives [38, 40, 62, 49, 69]. Although convolutional neural networks (CNNs) inherently learn the composition of visual parts [78, 5, 31], they do not explicitly tie visual representation and language semantics in a compositional way. Recently, several concurrent works [59, 56, 35] extend the pre-training strategies from BERT [9] to learn the latent compositional representations, which jointly represent images and descriptive texts for solving VQA, captioning, or image-text matching. However, these works mostly fix the image representation pre-extracted from a detection [51] or recognition [71] model. This not only limits their applicability to certain imagery domain, but also leads to an overall complex, heavy modelling framework. We propose a remedy by injecting language semantics at varying depths inside a CNN. This effectively yields a more powerful composite representation with simpler, lighter modelling.

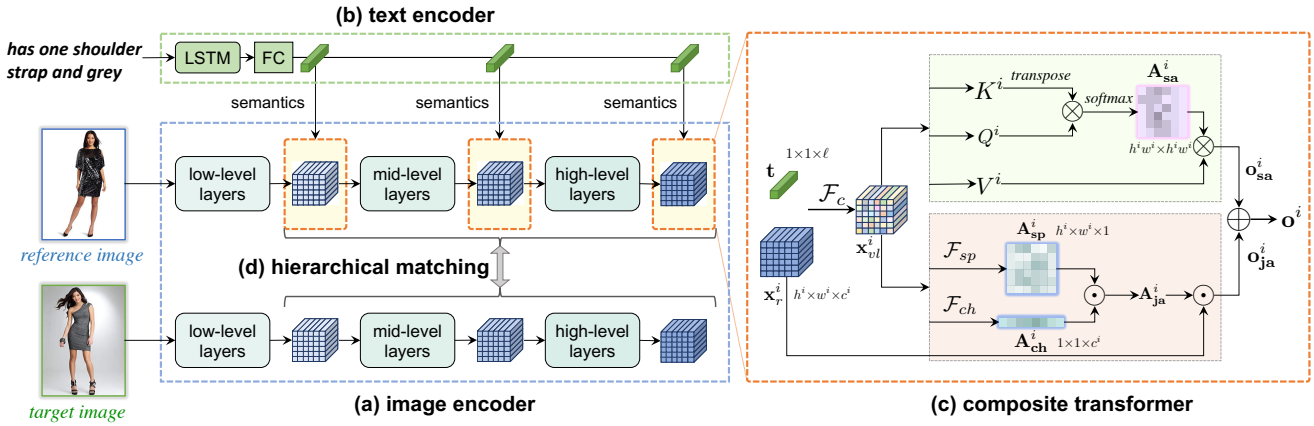


Figure 2. An overview of our **V**isiolinguistic **A**ttention **L**earning (**VAL**) framework. Given a pair of *reference image* and *text* as input, our goal is to learn their composite representation that aligns exclusively to the *target image* representation. VAL contains three major components: **(a)** an image encoder and **(b)** a text encoder (Sec. 3.1), **(c)** composite transformers (Sec. 3.2) that are plugged into different convolution layers to compose the vision and language contents. All components are synergistically optimised by **(d)** hierarchical matching (Sec. 3.3). Symbols of $\odot$, $\otimes$, $\oplus$ stand for the Hadamard product, matrix multiplication and element-wise addition, respectively.

3. Visiolinguistic Attention Learning

Fig. 2 presents an overview of our **V**isiolinguistic **A**ttention **L**earning (VAL) framework. Given a *reference image* and *user text* as input query, the ultimate aim of VAL is to learn a composite representation that aligns exclusively to the *target image* representation. VAL contains three components: **(a)** an image encoder, **(b)** a text encoder for vision and language representation learning; and **(c)** multiple composite transformers that absorb language semantics into visual feature maps at varying depths. All components are jointly optimised in an end-to-end manner via a hierarchical matching objective. We start with an overview of two basic components in Sec. 3.1, then elaborate our key ingredient and model optimisation in Sec. 3.2, Sec. 3.3.

3.1. Representing Images and Texts

Image Representation. To encapsulate the visual contents into discriminative representations, we employ an image encoder, i.e. a standard CNN, for image representation learning. As CNNs inherently learn visual concepts of increasing abstraction in a compositional, hierarchical order [5, 31, 78], we conjecture that image features from a single convolution layer do not capture the visual information of different granularities. Thus, we extract the feature maps from multiple convolution layers to construct a build-in feature pyramid [33] for more expressive representation learning. Concretely, the feature pyramid F is obtained from three different levels inside the CNN θ_{CNN}:

$$F_r = \{\mathbf{x}_r^L, \mathbf{x}_r^M, \mathbf{x}_r^H\} = \theta_{\text{CNN}}(\mathbf{I}_r)$$
$$F_t = \{\mathbf{x}_t^L, \mathbf{x}_t^M, \mathbf{x}_t^H\} = \theta_{\text{CNN}}(\mathbf{I}_t)$$

Here, $\mathbf{I}_r, \mathbf{I}_t$ refer to the *reference image* and *target image*; F_r, F_t are their corresponding feature pyramids, with each containing multi-level feature maps $\mathbf{x}^L, \mathbf{x}^M, \mathbf{x}^H$ extracted from the *Low*, *Mid*, *High*-level convolution layers[1].

Text Representation. To represent the semantics of texts, we utilise a text encoder to map the *user text* $\mathbf{T}$ into a vectorised text representation. Formally, the text encoder is implemented as an LSTM, followed by max-pooling and a linear projection layer. In brief, we first apply basic tokenising on text, then feed the token sequence into the text encoder to obtain the final text representation: $\mathbf{t} \in \mathbb{R}^\ell$.

3.2. Composite Transformer

To jointly represent images and texts, we propose to *transform* and *preserve* the visual features conditioned on language semantics. Inspired by the superiority of transformer [64] in multimodal learning [23, 35], we devise a composite transformer plugged at multi-level inside a CNN. Our key idea is to learn a composite representation of image and text through *attentional transformation* and *preservation* learnt upon the *visiolinguistic features* (Fig. 2(**c**)), with the ultimate aim to capsule the essential visual and linguistic contents for visual search, which we describe next.

Visiolinguistic Representation. To digest the information flows from vision and language domains, the reference image feature F_r, text feature $\mathbf{t}$ are first fused to obtain the visiolinguistic represention. Formally, for feature maps $\mathbf{x}_r^i$ (where i=L, M, H is the level in feature pyramid), multimodal fusion is performed by concatenation with the text feature $\mathbf{t}$, followed by a composite function $\mathcal{F}_c$ to learn the fused visiolinguistic feature $\mathbf{x}_{vl}^i$:

$$\mathbf{x}_{vl}^i = \mathcal{F}_c([\mathbf{x}_r^i, \mathbf{t}]) \quad (1)$$

[1] Refer to *Supplementary Material* for more architecture details.

where $[\cdot,\cdot]$ denotes concatenation, which broadcasts the text feature $\mathbf{t}$ spatially to match the shape of image feature $\mathbf{x}_r^i$; $\mathcal{F}_c$ is an MLP. Here, the input $\mathbf{x}_r^i$, output $\mathbf{x}_{vl}^i$ are kept as 3D feature tensors (i.e. $\mathbf{x}_r^i, \mathbf{x}_{vl}^i \in \mathbb{R}^{h^i \times w^i \times c^i}$) to ensure spatial information is not collapsed due to global pooling – each spatial vector conceptually corresponds to a part representation of image. Essentially, this composite process shares similar spirit as Relation Network [53], in that *pairwise visiolinguistic relationships* between the reference image and input text are formed spatially in the output $\mathbf{x}_{vl}^i$.

After fusing image and text features to the visiolinguistic feature $\mathbf{x}_{vl}^i$, we feed $\mathbf{x}_{vl}^i$ to a two-stream module for learning the *attentional transformation* and *preservation*.

Self-Attentional Transformation. To self-discover the *latent region-to-region relationships* essential for learning the transformation, we feed the visiolinguistic feature $\mathbf{x}_{vl}^i$ through a multi-head transformer[2]. The key insight is to capture the important visiolinguistic cues via *non-local self-attention* learning. This is achieved by first projecting $\mathbf{x}_{vl}^i$ into the latent space as query, key, value (i.e. Q, K, V):

$$Q^i = \mathcal{F}_Q(\mathbf{x}_{vl}^i),\ K^i = \mathcal{F}_K(\mathbf{x}_{vl}^i),\ V^i = \mathcal{F}_V(\mathbf{x}_{vl}^i)$$

where $\mathcal{F}_Q, \mathcal{F}_K, \mathcal{F}_V$ are implemented as 1×1 convolutions; $Q^i, K^i, V^i \in \mathbb{R}^{h^i \times w^i \times \bar{c}^i}$ are outputs in the latent space. The *self-attention* is then derived by reshaping Q^i, K^i to $\mathbb{R}^{n \times \bar{c}^i}$ ($n{=}h{\times}w$), followed with matrix multiplication:

$$\mathbf{A}_{\mathbf{sa}}^i = \text{softmax}(\frac{Q^i {K^i}^T}{\sqrt{\bar{c}}})$$

where $\mathbf{A}_{\mathbf{sa}}^i \in \mathbb{R}^{n \times n}$ is the *self-attention matrix*, with each element indicating the intensity of focus when learning the transformation. The output of this stream is updated by aggregating the essential information from the latent representation V, followed by a linear transformation layer $\mathcal{F}_{sa}$:

$$\mathbf{o}_{\mathbf{sa}}^i = \mathcal{F}_{sa}(\mathbf{A}_{\mathbf{sa}}^i V) \tag{2}$$

where $\mathbf{o}_{\mathbf{sa}}^i \in \mathbb{R}^{h^i \times w^i \times c^i}$. In essence, this self-attentional stream learns the *non-local interactions* [68, 50] among the *pairwise visiolinguistic relationships* formed in $\mathbf{x}_{vl}^i$. Per visiolinguistic relationship, it generates an attention mask to highlight the spatial long-range interdependencies that are essential for learning the feature transformation.

Joint-Attentional Preservation. Whilst self-attention captures the non-local correlations for feature transformation, it does not specify how should the reference image feature $\mathbf{x}_r^i$ be preserved to resemble the input image $\mathbf{I}_r$. To retain the unaltered visual content in $\mathbf{I}_r$, we introduce a joint-attentional stream alongside the self-attentional stream. Specifically, this stream contains spatial-channel attention learnt upon on the visiolinguistic feature $\mathbf{x}_{vl}^i$ to recalibrate the strength of preservation on $\mathbf{x}_r^i$. This is motivated that different feature maps encode different semantics, e.g. colors, materials, parts [80]. Thus, to selectively suppress and highlight the visual content in $\mathbf{I}_r$, attentional preservation is introduced to selectively reuse the reference image feature $\mathbf{x}_r^i$. Formally, a *lightweight joint-attention* is learnt upon on the visiolinguistic feature $\mathbf{x}_{vl}^i$ in a squeeze-and-excite manner [22] to obtain the selective activation on $\mathbf{x}_r^i$:

$$\mathbf{A}_{\mathbf{sp}}^i = \text{sigmoid}(\mathcal{F}_{sp}(\frac{1}{c^i}\sum_j^{c^i} \mathbf{x}_{vl}^i(:,:,j)))$$

$$\mathbf{A}_{\mathbf{ch}}^i = \text{sigmoid}(\mathcal{F}_{ch}(\frac{1}{h^i{\times}w^i}\sum_j^{h^i}\sum_k^{w^i} \mathbf{x}_{vl}^i(j,k,:)))$$

$$\mathbf{A}_{\mathbf{ja}}^i = \mathbf{A}_{\mathbf{sp}}^i \odot \mathbf{A}_{\mathbf{ch}}^i$$

where $\mathbf{A}_{\mathbf{sp}}^i{\in}\mathbb{R}^{h^i \times w^i \times 1}$, $\mathbf{A}_{\mathbf{ch}}^i{\in}\mathbb{R}^{1\times1\times c^i}$, $\mathbf{A}_{\mathbf{ja}}^i{\in}\mathbb{R}^{h^i \times w^i \times c^i}$; $\mathcal{F}_{sp}$, $\mathcal{F}_{ch}$ are implemented as $h^i{\times}w^i$, $1{\times}1$ convolutions to learn the spatial, channel attentions $\mathbf{A}_{\mathbf{sp}}^i, \mathbf{A}_{\mathbf{ch}}^i$. $\mathbf{A}_{\mathbf{ja}}^i$ is the *joint-attention matrix* derived from $\mathbf{A}_{\mathbf{sp}}^i, \mathbf{A}_{\mathbf{ch}}^i$, which dynamically modulates the intensity to preserve the reference image feature $\mathbf{x}_r^i$:

$$\mathbf{o}_{\mathbf{ja}}^i = \mathbf{A}_{\mathbf{ja}}^i \odot \mathbf{x}_r^i \tag{3}$$

where $\mathbf{o}_{\mathbf{ja}}^i \in \mathbb{R}^{h^i \times w^i \times c^i}$. The final output of the composite transformer is the weighted sum of outputs $\mathbf{o}_{\mathbf{sa}}^i$, $\mathbf{o}_{\mathbf{ja}}^i$ from two complementary attentional streams:

$$\mathbf{o}^i = w_{sa}\mathbf{o}_{\mathbf{sa}}^i + w_{ja}\mathbf{o}_{\mathbf{ja}}^i \tag{4}$$

where w_{sa}, w_{ja} are learnable scalars to control the relative importance of two streams. The composite output of VAL is denoted as $F_o = \{\mathbf{o}^L, \mathbf{o}^M, \mathbf{o}^H\}$ – a feature pyramid with each level derived from one composite transformer. The final composite feature used for image retrieval is simply the concatenation of multi-level outputs after average-pooling.

3.3. Hierarchical Matching

As our ultimate aim is to align the composite output F_o and the target image representation F_t exclusively, we formulate a hierarchical matching objective, with two losses formed in a two-level hierarchy to match with the desired *visual* and *semantic* features (Fig. 3), as detailed next.

Primary visual-visual matching. We introduce visual-visual matching as our *primary objective* to ensure the composite feature match the target feature with high similarity. Formally, with similarity measured by L2 distance d, a bi-directional triplet ranking loss [10] is imposed to align the multi-level feature maps in two feature pyramids F_o, F_t:

$$\mathcal{L}_{vv} = \sum_i^{L,M,H} \underbrace{\mathcal{L}_i(\bar{\mathbf{o}}^i, \bar{\mathbf{x}}_t^i)}_{\text{rank } \bar{\mathbf{x}}} + \underbrace{\mathcal{L}_i(\bar{\mathbf{x}}_t^i, \bar{\mathbf{o}}^i)}_{\text{rank } \bar{\mathbf{o}}} \tag{5}$$

$$\text{with } \mathcal{L}_i(\bar{\mathbf{o}}, \bar{\mathbf{x}}_t^i) = \max(0, d(\bar{\mathbf{o}}^i, \bar{\mathbf{x}}_t^i) - d(\bar{\mathbf{o}}^i, \bar{\mathbf{x}}_n^i) + m)$$

[2] We omit the multi-head formulation [64] of tensor split and concatenation to avoid clutter. Details are given in *Supplementary Material*.

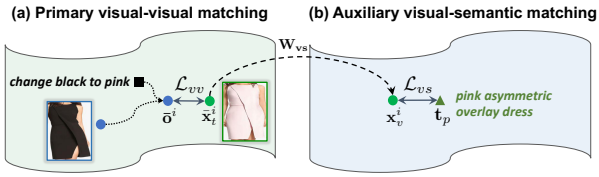


Figure 3. Feature matching in a two-level hierarchical space.

Here, $\bar{\mathbf{o}}^i, \bar{\mathbf{x}}_t^i$ are average-pooled features at i_{th} level in feature pyramids F_o, F_t; m is distance margin. We adopt *semi-hard mining* [54] to select the negative pair $\bar{\mathbf{x}}_n^i$. $\mathcal{L}_{vv}$ constrains attention learning at multi-level to incentivise multi-granular alignments across the network. Per level, $\mathcal{L}_i$ encourages the composite feature $\bar{\mathbf{o}}^i$ to match the target image feature $\bar{\mathbf{x}}_t^i$ with a smaller distance than the negative pair $\bar{\mathbf{x}}_n^i$.

Auxiliary visual-semantic matching. To further tie the learnt representation with desired semantics, we introduce visual-semantic matching as an *auxiliary regulariser*. This is beneficial when images are tagged with descriptive texts (e.g. product descriptions) to serve as side information during training [55, 32]. Formally, a bi-directional triplet ranking loss is imposed to align the projected visual feature and its corresponding text feature in a shared embedding space (Fig. 3**(b)**:

$$\mathcal{L}_{vs} = \sum_{i}^{L,M,H} \underbrace{\mathcal{L}_i(\mathbf{x}_v^i, \mathbf{t}_p)}_{\text{rank } \mathbf{t}} + \underbrace{\mathcal{L}_i(\mathbf{t}_p, \mathbf{x}_v^i)}_{\text{rank } \mathbf{x}}$$
$$\text{with } \mathcal{L}_i(\mathbf{x}_v^i, \mathbf{t}_p) = \max(0, d(\mathbf{x}_v^i, \mathbf{t}_p) - d(\mathbf{x}_v^i, \mathbf{t}_n) + m) \tag{6}$$

Here, $\mathbf{x}_v^i \in \mathbb{R}^\ell$ is the projected visual feature mapped from the visual space to the semantic space by a linear projection $\mathbf{W_{vs}}$; $\mathbf{t}_p$, $\mathbf{t}_n$ are positive, negative text pairs. $\mathcal{L}_{vs}$ essentially acts as a regulariser by aligning the projected feature and its text feature, which can be imposed via *pre-training* or *joint training* with Eq. 5 to tie visual representations with corresponding semantics in a meaningful way.

4. Experiments

4.1. Experimental Setup

Datasets. To validate the model's generalisability to *various text feedbacks*, we evaluate on three datasets, including **(1)** Fashion200k using *attribute-like description*, **(2)** Shoes and FashionIQ using *natural language expression*. We details these datasets in Sec. 4.2, Sec. 4.3 and Sec. 4.4.

Compared Methods. To validate the efficacy of our approach in image search with text feedback, we compare with four representative multimodal learning methods:

- Relationship [53]: A relation reasoning module. It takes in feature maps extracted from the final layer of a CNN and text feature from an RNN, followed by concatenation and an MLP to learn the cross-modal relationships. The pairwise relationships are simply summed and processed through another MLP to get the final output.
- FiLM [47]: A Feature-wise Linear Modulation component. It contains a stack of three FiLM layers cascaded after a CNN. The text information is represented by the text feature extracted from an RNN to modulate each feature map by affine transformation.
- MRN [25]: A Multimodal Residual Learning component. It learns multimodal representations by fusing visual and textual features from a CNN and an RNN. The cross-modal features are obtained through three blocks of element-wise multiplication and residual learning.
- TIRG [66]: An image-text composition approach for image retrieval. It composes visual and textual features by concatenation, followed by learning a gating connection and a residual connection for cross-modal fusion.

Discussion. Among the above methods, TIRG is proposed for image search with attribute-like text feedback; whilst others are originally used in VQA. However, unlike existing methods that stack *transformation layers* after a CNN, VAL uniquely plugs the *composite transformers* at multi-level inside a CNN to capture multi-granular visiolinguistic information. In addition, VAL is specially featured with *two attentional streams* that operate upon the visiolinguistic features to selectively *transform* and *preserve* the visual features conditioned on the language semantics. For a fair comparison, we implement existing methods using the same CNN, RNN trained by a bi-directional ranking loss.

Ablative baselines. Besides comparing with existing methods, we conduct several ablative tests on our model:

- **VAL** ($\mathcal{L}_{vv}$): VAL optimised with the primary objective (Eq. 5), i.e. auxiliary regulariser (Eq. 6) is not used.
- **VAL** ($\mathcal{L}_{vv}+\mathcal{L}_{vs}$): VAL trained by hierarchical matching, using side information by joint training or pre-training.
- **VAL** (GloVe): It shares the same structure as VAL ($\mathcal{L}_{vv}+\mathcal{L}_{vs}$), with word vectors initialised from GloVe [46].

The latter two tests endow our VAL model with prior linguistic knowledge from side information and GloVe.

Implementation Details. We conduct all the experiments in Tensorflow [1]. We initialise the CNNs pre-trained from ImageNet [8], and integrate the composite transformers into ResNet-50 [19] on Shoes, FashionIQ, and MobileNet [21] on Fashion200k. In the self-attentional stream, we set the number of heads to 2. The LSTM [20] is one-layer with 1024 hidden units, followed by a linear projection layer that maps the max-pooled LSTM feature to the text feature of 512 dimension. We use Adam [26] optimiser with a constant learning rate of 2×10^{-4} and α, β of 0.999, 1×10^{-8}. The batch size is set to 32. The margin m in Eq. 5, Eq. 6 is set to 0.2. More network architecture and training details are given in *Supplementary Material* due to space limit.

Evaluation Metric. We adopt the standard evaluation metric in retrieval, i.e. Recall@K, denoted as R@K for short.

Figure 4. Qualitative results of image search with *attribute-like* text feedback on Fashion200k. blue/green boxes: reference/target images.

Method	R@1	R@10	R@50
Han et al. [18]	6.3	19.9	38.3
Show and Tell [65]	12.3	40.2	61.8
Param Hashing [43]	12.2	40.0	61.7
FiLM [47]	12.9	39.5	61.9
Relationship [53]	13.0	40.5	62.4
MRN [25]	13.4	40.0	61.9
TIRG [66]	14.1	42.5	63.8
MRN	14.2	43.6	63.8
TIRG	14.8	43.7	64.1
VAL ($\mathcal{L}_{vv}$)	**21.2**	**49.0**	**68.8**
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$)	**21.5**	**53.8**	**73.3**
VAL (GloVe)	**22.9**	**50.8**	**72.7**

Table 1. Quantitative results of image search with text feedback on Fashion200k. Rows in colours indicate results obtained with the *same* networks and data. Overall $1^{st}/2^{nd}$ best in **red**/**blue**.

4.2. Fashion200k

Fashion200k [18] 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. It also covers a diverse range of fashion concepts, with a total vocabulary size of 5,590. Each image is tagged with descriptive texts as product description, such as "`white logo print t-shirt`", which is exploited as side information for auxiliary supervision via joint training. Following [66], we use the training split of around 172k 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 (see *Supplementary Material*).

Table 1 shows our comparison with existing methods. We reproduce the best competitors with the same networks and optimiser for a like-to-like fair comparison. As can be seen, our model demonstrates compelling results compared to all other alternatives, e.g. VAL ($\mathcal{L}_{vv}$) outperforms the best competitor TIRG with an improved margin of 6.4% in R@1. We also observe that **(1)** VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$) performs better than VAL ($\mathcal{L}_{vv}$), which indicates the advantage of introducing auxiliary supervision to match with additional semantics; **(2)** VAL (GloVe) performs on par with VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$), suggesting using GloVe word vectors is not so vital when using attribute-like text feedback.

Method	R@1	R@10	R@50
FiLM	10.19	38.89	68.30
MRN	11.74	41.70	67.01
Relationship	12.31	45.10	71.45
TIRG	12.60	45.45	69.39
VAL ($\mathcal{L}_{vv}$)	**16.49**	**49.12**	**73.53**
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$)	**16.98**	**49.83**	**73.91**
VAL (GloVe)	**17.18**	**51.52**	**75.83**

Table 2. Quantitative results of image search with text feedback on Shoes. Rows in colour indicate results obtained with the *same* networks and data. Overall $1^{st}/2^{nd}$ best in **red**/**blue**.

Fig. 4 shows our qualitative results on Fashion200k. We notice our model is able to retrieve new images that resemble the reference image, while changing certain attributes conditioned on text feedback, e.g. colour, material and trim.

4.3. Shoes

Shoes [6] is a dataset originally crawled from *like.com*. It is further tagged with *relative captions in natural language* for dialog-based interactive retrieval [16]. Following [16], we use 10,000 training samples for training and 4,658 test samples for evaluation. Besides relative captions, there are 3,000 images tagged with descriptive texts, such as "`brown buckle mules`", which are used as auxiliary supervision (Eq. 6) for pre-training in VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$). Due to missing results of state-of-the-art methods in composing image and text for image search, we provide a new benchmark on this dataset by performing experiments under the same networks and optimiser for a comprehensive comparison.

Table 4 shows the clear superiority of our model compared to other alternatives. For instance, VAL ($\mathcal{L}_{vv}$) surpasses the best competitor TIRG by 3.89% in R@1. We also notice the clear advantages of utilising prior linguistic knowledge in VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$) and VAL (GloVe), as compared to not using such knowledge in VAL ($\mathcal{L}_{vv}$).

Fig. 5 further shows our qualitative results on Shoes. It suggests our model is capable of ingesting multiple visual attributes and properties in the natural language text feedback to search for the desired target images. More qualitative results are given in *Supplementary Material*.

Figure 5. Qualitative results of image search with *natural language* text feedback on Shoes. blue/green boxes: reference/target images.

Method	Dress		Shirt		Toptee		Avg	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
TIRG	8.10	23.27	11.06	28.08	7.71	23.44	8.96	24.93
Image+Text Concatenation	10.52	28.98	13.44	34.60	11.36	30.42	11.77	31.33
Side Information [17]	11.24	32.39	13.73	37.03	13.52	34.73	12.82	34.72
MRN	12.32	32.18	15.88	34.33	18.11	36.33	15.44	34.28
FiLM	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04
TIRG	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39
Relationship	15.44	38.08	18.33	38.63	21.10	44.77	18.29	40.49
VAL ($\mathcal{L}_{vv}$)	**21.12**	**42.19**	**21.03**	**43.44**	**25.64**	**49.49**	**22.60**	**45.04**
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$)	**21.47**	**43.83**	**21.03**	**42.75**	**26.71**	**51.81**	**23.07**	**46.13**
VAL (GloVe)	**22.53**	**44.00**	**22.38**	**44.15**	**27.53**	**51.68**	**24.15**	**46.61**

Table 3. Quantitative results of image search with text feedback on FashionIQ. Avg: averaged R@10/50 computed over three categories. Rows in colour indicate results obtained with the *same* backbone networks (i.e. CNN, LSTM) and data. Overall $1^{st}/2^{nd}$ best in **red**/**blue**.

Figure 6. Qualitative results of image search with *natural language* text feedback on FashionIQ. blue/green boxes: reference/target images.

4.4. FashionIQ

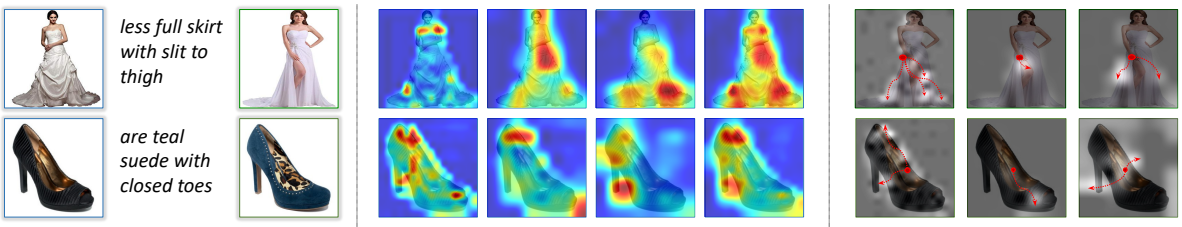
FashionIQ [17] is a *natural language* based interactive fashion product retrieval dataset. It contains 77,684 images crawled from *Amazon.com*, covering three categories: `Dresses`, `Tops&Tees` and `Shirts`. Among the 46,609 training images, there are 18,000 image pairs, with each pair accompanied with around two natural language sentences that describe one or multiple visual properties to modify in the reference image, such as "`is darker`" and "`has short sleeves and is longer and more flowing`". We use the side information from Fashion200k as auxiliary supervision for pre-training in VAL ($\mathcal{L}_{vv}+\mathcal{L}_{vs}$). Following the same evaluation protocol of composing image and text for retrieval [17], we use the same training split and evaluate on the validation set[3]. We report results on individual category, as well as the averaged results over three categories[4].

Table 3 shows our model outperforms other competitors substantially, e.g. VAL ($\mathcal{L}_{vv}$) surpasses Relationship with an overall margin of 4.31% in R@10. We also notice the performance boosts in VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$) and VAL (GloVe), as compared to VAL ($\mathcal{L}_{vv}$). This again indicates the benefit of using prior linguistic knowledge from auxiliary semantics and GloVe when using natural language text feedback.

Fig. 6 presents our qualitative results on FashionIQ. It shows that given multiple semantic concepts within a sentence snippet, our model captures both concrete and abstract semantics, including various fashion elements [63]

[3]The groundtruth of test set in FashionIQ has not been released yet.

[4]The unpublished state-of-the-art uses an ensemble of diverse models.



(a) Example pairs **(b)** Joint-attention **(c)** Self-attention

Figure 7. Attention visualisation. **(a)** Example pairs of reference image, user text as input query, and desired target image output. **(b)** The attended regions (i.e. the maximum magnitude along the channel dimension) for preservation selected by joint-attention at multi-levels; and average attended regions across all levels. **(c)** The self-attention at the central query point, with arrows indicating the attended regions.

Method	Fashion200k		FashionIQ (Avg)		Shoes	
	R@1	R@10	R@10	R@50	R@1	R@10
w/o SA	16.3	46.9	21.94	44.56	7.85	42.33
w/o JA	19.9	48.8	21.31	43.74	13.43	42.01
VAL	**21.2**	**49.0**	**22.60**	**45.04**	**16.49**	**50.09**

Table 4. Ablation study on effect of attention learning.

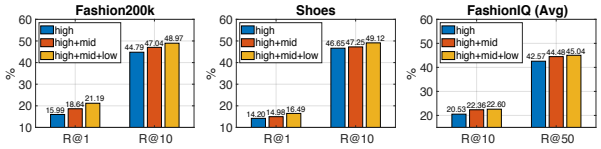


Figure 8. Effect of composition at multi-level.

like colour, silhouette, printing, etc. We also observe that our model can jointly comprehend the global appearance (e.g. overall colours, patterns), as well as local fine-grained details (e.g. a specific logo and trim) for image search.

4.5. Ablation Study

In this section, we conduct analysis to give an insight of the *key ingredient* in VAL (i.e. composite transformers). We perform experiments with the primary objective (Eq. 5) to exclude the effect of auxiliary regulariser.

Effect of self-attention and joint-attention. To analyse the synergistic effect of *self-attentional transformation* (SA) and *joint-attention preservation* (JA), we compare our composite transformer with two baselines: **(a)** remove SA stream (i.e. "w/o SA"); **(b)** remove JA stream (i.e. "w/o JA") – see a graphical illustration in *Supplementary Material*. For each baseline, we remove one attentional stream to study its effect. Table 4 shows the comparison on FashionIQ and Shoes. It can be seen that our VAL does profit substantially from the complementary benefits of SA and JA. This verifies our rationale of composing visual features and language semantics through *attentional transformation* and *preservation* learnt upon the visiolinguistic features.

Attention visualisation. To further interpret the attentions learnt by VAL at varying representation depths (i.e. low, mid, high level), we visualise the attended regions by joint-attention and self-attention in Fig. 7. From Fig. 7**(b)**, we notice that the spatially attended region varies across different levels. This indicates the joint-attention stream picks up different visual cues to preserve across varying depths. From Fig. 7**(c)**, we observe that the multi-level self-attention trigger various attended regions for learning the transformation, e.g. in the *dress* example, the low-level self-attention highlights the overall silhouette, while the mid, high-level self-attentions pick up the `thigh` area to focus on.

Overall, Fig. 7 shows our model captures visual cues at different granularities to selectively *preserve* and *transform* the reference image features according to language semantics. This suggests that VAL learns to capture the essential multi-granular visiolinguistic contents for image search.

Effect of composition at multi-level. We test how composition at multi-level aids in representation learning by comparing VAL (high+mid+low) to two baselines: **(a)** high, **(b)** high+mid, which perform composition at high or high+mid level. Fig. 8 shows composition at multi-level improves the overall performance. This verifies the efficacy of employing composite transformers at varying depths to capture the multi-granular information, which also accords with the fact that CNNs learn visual features of increasing abstraction from lower to higher layers [58]. While focusing on multi-modal representation learning, our model can also be integrated with a dialogue manager [16] for interactive search.

5. Conclusion

We introduced VAL, a novel approach to tackle the challenging task of image search with text feedback. VAL is featured with multiple *composite transformers* that selectively *preserve* and *transform* multi-level visual features conditioned on semantics to derive an expressive composite representation. We validate the efficacy of VAL on three datasets, and demonstrate its consistent superiority in handling various text feedbacks, including *attribute-like description* and *natural language expression*. We also explore *auxiliary semantics* to further boost the model performance. Overall, this work provides a novel approach along with a comprehensive evaluation, which collectively advance the research in interactive visual search using text feedback.

Acknowledgement: We would like to thank Maksim Lapin, Michael Donoser, Bojan Pepik, and Sabine Sternig for their helpful discussions.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016. 5

[2] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[5] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2009. 2, 3

[6] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, 2010. 6

[7] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 2002. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics*, 2019. 2

[10] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*, 2018. 4

[11] Marin Ferecatu and Donald Geman. Interactive search for image categories by mental matching. In *IEEE International Conference on Computer Vision*, 2007. 1

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, 2013. 1, 2

[13] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[14] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *IEEE International Conference on Computer Vision*, 2019. 2

[15] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, 2016. 1

[16] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, 2018. 1, 2, 6, 8

[17] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019. 1, 2, 7

[18] 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 *IEEE International Conference on Computer Vision*, 2017. 1, 2, 6

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 5

[21] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[23] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *IEEE International Conference on Computer Vision*, 2019. 2, 3

[24] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[25] 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. 5, 6

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1

[28] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2

[29] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015. 2

[30] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 2017. 2

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 2015. 2, 3

[32] Wen Li, Li Niu, and Dong Xu. Exploiting privileged information from web data for image categorization. In *European Conference on Computer Vision*, 2014. 5

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[34] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[35] 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. 2, 3

[36] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, 2016. 2

[37] Long Mai, Hailin Jin, Zhe Lin, Chen Fang, Jonathan Brandt, and Feng Liu. Spatial-semantic image search by visual feature synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[38] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[39] Nils Murrugarra-Llerena and Adriana Kovashka. Image retrieval with mixed initiative and multimodal feedback. In *British Machine Vision Conference*, 2018. 1

[40] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *European Conference on Computer Vision*, 2018. 2

[41] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, 2018. 2

[42] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *IEEE International Conference on Computer Vision*, 2017. 1

[43] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6

[44] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[45] Devi Parikh and Kristen Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, 2011. 1, 2

[46] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014. 5

[47] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. 5, 6

[48] Bryan Plummer, Hadi Kiapour, Shuai Zheng, and Robinson Piramuthu. Give me a hint! navigating image databases using human-in-the-loop feedback. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 1

[49] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *IEEE International Conference on Computer Vision*, 2019. 2

[50] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, 2019. 4

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 2

[52] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on circuits and systems for video technology*, 1998. 1

[53] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, 2017. 4, 5, 6

[54] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 5

[55] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *IEEE International Conference on Computer Vision*, 2013. 5

[56] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *IEEE International Conference on Computer Vision*, 2019. 2

[57] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 8

[59] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 2

[60] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[61] Bart Thomee and Michael S Lew. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval*, 2012. 1

[62] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *IEEE International Conference on Computer Vision*, 2019. 2

[63] Kristen Vaccaro, Sunaya Shivakumar, Ziqiao Ding, Karrie Karahalios, and Ranjitha Kumar. The elements of fashion style. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016. 7

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2, 3, 4

[65] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2, 6

[66] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 5, 6

[67] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2

[68] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[69] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *IEEE International Conference on Computer Vision*, 2019. 2

[70] Hong Wu, Hanqing Lu, and Songde Ma. Willhunter: interactive image retrieval with multilevel relevance. In *IEEE International Conference on Pattern Recognition*, 2004. 1

[71] S Xie, C Sun, J Huang, Z Tu, and K Murphy. Rethinking spatiotemporal feature learning for video understanding (2017). arxiv preprint. In *European Conference on Computer Vision*, 2018. 2

[72] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, 2016. 2

[73] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[74] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *European Conference on Computer Vision*, 2018. 2

[75] Aron Yu and Kristen Grauman. Thinking outside the pool: Active training image creation for relative attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[76] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[77] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Advances in Neural Information Processing Systems*, 2019. 2

[78] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer. 2, 3

[79] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[80] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 4

[81] Xiang Sean Zhou and Thomas S Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*, 2003. 1

[82] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured attentions for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

Image Search with Text Feedback by Visiolinguistic Attention Learning (Supplementary Material)

Yanbei Chen
Queen Mary University of London
yanbei.chen@qmul.ac.uk

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Loris Bazzani
Amazon
bazzanil@amazon.com

1. Architecture Details

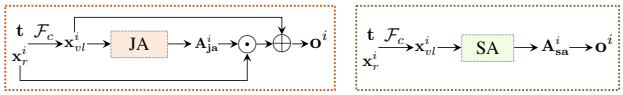
Part Name	Layer Description
composite function	$\mathcal{F}_c$: conv(K-1×1, N-c^i, ReLU)
SA stream	$\mathcal{F}_Q, \mathcal{F}_K, \mathcal{F}_V$: conv(K-1×1, N-$\bar{c}^i$)
	$\mathcal{F}_{sa}$: conv(K-1×1, N-c^i, ReLU)
JA stream	$\mathcal{F}_{sp}$: conv(K-h^i×w^i, N-1, sigmoid)
	$\mathcal{F}_{ch}$: conv(K-1×1, N-c^i, sigmoid)

Table 1. Architecture of composite transformer. SA: self-attention. JA: joint-attention. conv(K,N): stands for convolution layer, where K: filter size, N: number of filters. $\bar{c}^i = \frac{c^i}{\text{num\_heads}}$, num_heads $= 2$.

Training. Table 1 details the architecture of our composite transformer. To learn the transformation at varying depths, we plug three composite transformers into the CNN at the low, mid, high-level. In ResNet-50, the low, mid, high-level feature maps are from the last three residual blocks, which give feature tensors of size $16\times16\times512$, $8\times8\times1024$, $8\times8\times2048$. In MobileNet, the low, mid, high-level are set as the $6, 11, 13$th layer, which give feature tensors of size $16\times16\times512$, $16\times16\times512$, $8\times8\times1024$.

Testing. At inference time, outputs from three composite transformers are average-pooled and concatenated to derive the *composite feature*. The *test image feature* is simply the concatenation of average-pooled features at the low, mid, high-level. For retrieval, the *composite feature* is compared with *test image features* by measuring their pairwise similarities, formally computed as L2 distance.

Computational costs. At test time, the computational costs are decided by **(1)** model complexity (FLOPs); **(2)** matching and ranking. On **(1)**, our composite transformers have FLOPs (8.10×10^7) vs. ResNet50 (3.80×10^9), which bring small computational cost - an additional FLOPs of 2.13%. On **(2)**, the complexity is $\mathcal{O}(QN)$, $\mathcal{O}(QNlogN)$ for similarity matching, ranking – Q, N are the size of query, test set. We implemented similarity matching on GPU, which yields lower computational cost than CPU implementation.



(a) w/o $\mathbf{A}^i_{\mathbf{sa}}$ **(b)** w/o $\mathbf{A}^i_{\mathbf{ja}}$

Figure 1. A schematic illustration of two baselines.

Ablative Baselines. As aforementioned, we test our VAL in comparison to two ablative baselines: **(a)** w/o self-attention (w/o $\mathbf{A}^i_{\mathbf{sa}}$), and **(b)** w/o joint-attention (w/o $\mathbf{A}^i_{\mathbf{ja}}$). We show a graphical illustration of these two baselines in Fig. 1.

2. Dataset and Training Details

(a) **(b)**

Figure 2. Example image pairs on Fashion200k. For each pair, we show the tagged attribute-like product descriptions of the *reference image* and *target image*; while the *user text* (in quotation marks) describes the difference between two images in *attributes*.

(a) **(b)**

Figure 3. Examples on Shoes. **(a)** Image pair with *relative caption* in natural language. **(b)** Example image with tagged description.

Figure 4. Image pair tagged with *relative captions* on FashionIQ.

We present illustration of different datasets in Fig. 2, 3 and 4. Below, we detail how each dataset is used in training.

Fashion200k. In training, we utilise the samples that can find their corresponding pairs with one word difference by comparing the tagged attribute-like product descriptions, as shown in Fig. 2. In VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$), we exploit the tagged

Figure 5. Qualitative results on Shoes. First two examples: "***success***" cases with small R@K (i.e. R@1, R@2); Last two examples: "***failure***" cases with relatively larger R@K (i.e. R@6, R@10). blue/green boxes: reference/target images.

Figure 6. Qualitative results on FashionIQ. First two examples: "***success***" cases with small R@K (i.e. R@1, R@2); Last two examples: "***failure***" cases with relatively larger R@K (i.e. R@5, R@10). blue/green boxes: reference/target images.

descriptions as side information, which serve as auxiliary supervision to train our VAL via a *joint-training* objective as $\mathcal{L}_{vv} + \mathcal{L}_{vs}$. We train for 160k iterations on Fashion200k.

Shoes. In training, we use 17,954 image pairs with relative captions (Fig. 3**(a)**). In VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$), we use the tagged descriptions (Fig. 3**(b)**) of 3,000 samples as side information for *pre-training* via $\mathcal{L}_{vs}$; then we fine-tune with the primary objective $\mathcal{L}_{vv}$ for 30 iterations.

FashionIQ. In training, we use all training pairs with tagged relative captions (Fig 4). In VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$), due to missing narrative descriptive texts tagged for each sample, we exploit Fashion200k to provide auxiliary supervision ($\mathcal{L}_{vs}$) via *pre-training*, and fine-tune with the primary objective $\mathcal{L}_{vv}$. We train for 50k iterations on FashionIQ.

3. Additional Qualitative Results

We provide additional qualitative results on Shoes and FashionIQ (Fig. 5, 6) to further give an in-depth analysis when using *natural language* based text feedback.

Further Analysis. Unlike attribute-like text feedback that generally describes a concrete visual concept, *natural language* text feedback may be highly *abstract*, thus likely to be *ambiguous* and indicate *multiple possibilities*. As Fig. 5, 6 show, there are multiple "*failure*" cases, which show the model does properly return the "desired" images that resemble the reference images whilst reflecting changes specified in the input texts. For instance, in the 3rd example in Fig. 6, there are more than one dress that contains "`a light floral pattern`" in the top retrieved items; but the target image is only R@5, mostly because the input text does indicate multiple possible desired outcomes.

Future Work. Overall, these results suggest that natural language based text feedback could sometimes be *ambiguous*, and thus indicate *multiple possible* desired items rather than a single one. To further examine or address this issue, we consider there are several potential research directions: (1) propose *new evaluation metrics* to quantify visual similarities among the top retrieved items; and (2) conduct *human studies* to test the interactive image retrieval performance in practical use.