FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback

Sonam Goenka[*†], Zhaoheng Zheng[*‡], Ayush Jaiswal[†],
Rakesh Chada[†], Yue Wu[†], Varsha Hedau[†], Pradeep Natarajan[†]
[†]Amazon Alexa Natural Understanding, [‡] USC Viterbi School of Engineering
[†]{goenkasg, ayujaisw, rakchada, wuayue, hedau, natarap}@amazon.com, [‡]zhaoheng.zheng@usc.edu

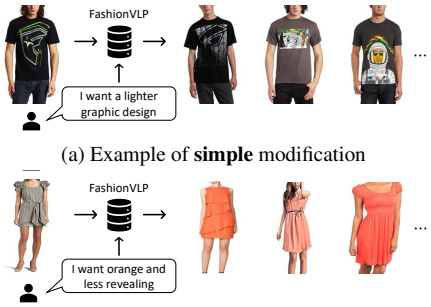
Abstract

Fashion image retrieval based on a query pair of reference image and natural language feedback is a challenging task that requires models to assess fashion related information from visual and textual modalities simultaneously. We propose a new vision-language transformer based model, FashionVLP, that brings the prior knowledge contained in large image-text corpora to the domain of fashion image retrieval, and combines visual information from multiple levels of context to effectively capture fashion-related information. While queries are encoded through the transformer layers, our asymmetric design adopts a novel attention-based approach for fusing target image features without involving text or transformer layers in the process. Extensive results show that FashionVLP achieves the state-of-the-art performance on benchmark datasets, with a large 23% relative improvement on the challenging FashionIQ dataset, which contains complex natural language feedback.

1. Introduction

The task of feedback-based fashion image retrieval involves fetching images of clothing items that match a customer's needs and preferences. A customer starts with an initial request to search for a fashion item and participates in multiple turns of interaction with the conversational assistant until they get the result that they are satisfied with. A key challenge in this use-case is to retrieve a new candidate image based on both the previously retrieved image and the new feedback provided by the customer. Figure 1 shows examples of feedback-based fashion image retrieval.

Substantial progress [6, 19, 25, 49] has been made on this topic by designing strong image-text composers using image and text features from separate neural networks. Recently, Vision-Language Pre-trained (VLP) transformers [8, 26, 28, 33, 44, 45, 54, 57, 59] have been shown to be capable of learning joint representations for images and text directly by training on large-scale image-text corpora. In this work, we propose a VLP transformer-based model, FashionVLP, for fashion image retrieval with textual feedback, which leverages prior knowledge from large corpora and image features from multiple fashion-related context levels.



(a) Example of **simple** modification

(b) Example of **compound** modification

Figure 1. Fashion image retrieval with textual feedback. The input query to the system includes a reference image and a comment specifying changes to be made to the image. The system retrieves fashion items with the desired changes accordingly.

Our model is composed of two parallel blocks – one for processing the reference image and the feedback, and another for processing target images. The reference block starts with extracting image features at multiple-levels of context: (1) whole image, (2) cropped image of clothing, (3) regions around fashion landmarks [32], and (4) regions of interest determined by a pretrained object detector. These features along with object tags from the detector and word tokens from the textual feedback are then fed into a multi-layer transformer model to compute a final joint representation for reference. On the target side, features at contexts (1)–(3) are computed using only image feature extractors for efficient low-cost inference, and fused using a contextual attention mechanism instead of transformer layers to generate a target encoding for each candidate image. The model is trained using cosine similarity and a batch-based classification loss where the target for each reference image is used as a negative sample for other reference images. Finally, retrieval is performed by ranking candidates using cosine similarities between reference and target encodings.

[*]Equal contribution. Completed during Zhaoheng's internship at Amazon.

We evaluate FashionVLP on three common fashion image retrieval datasets: FashionIQ [51], Shoes [3] and Fashion200K [17]. Unlike other datasets, FashionIQ contains real human comments on specific reference-target image pairs and is hence much more challenging for the fashion image retrieval task. Results show that FashionVLP improves the performance on FashionIQ by a significant relative performance gain of 23%. This validates the capability of our framework in dealing with complicated real-life image-feedback pairs when conducting fashion image retrieval. Our model also surpasses the state-of-the-art on Shoes and Fashion200K datasets.

Our work makes the following contributions:

- We propose a new transformer-based model that leverages prior knowledge from large image-text corpora for fashion image retrieval with textual feedback.
- We provide a way for effectively incorporating multiple levels of fashion-related visual context for both reference and candidate images within our asymmetric design.
- Our model outperforms previous works on benchmark datasets, with 23% relative gain on FashionIQ.

2. Related Work

Fashion Image Retrieval with Textual Feedback: The classic image retrieval task is a long-standing fundamental problem [9, 47] in computer vision, which requires comparison of reference and target images in a scalable way. Tremendous advances have been made in this field recently with the development of deep learning based methods [1, 15, 35, 37]. Alternative formulations use natural language text-based queries for image retrieval [10, 34, 41, 53, 55, 56].

Fashion image retrieval with textual feedback is different from the classic image retrieval problem as it takes both a reference image and a textual feedback for modifying the reference as query inputs, as shown in Fig. 1. Intuitively, this task can be solved via text-based visual relationship reasoning [21, 36, 40], where text features are injected into image feature extractors to get modified image encodings, which are then used for retrieval [36]. However, these methods do not explicitly combine visual and textual features into a joint semantic space, leading to poor performance.

In contrast, previous methods developed specifically for this task typically fuse the image and textual inputs into joint embeddings for retrieval. For example, TIRG [49] learns a gated feature and a residual feature for each image-text query and composes them into a joint encoding. The CIRPLANT [31] model fuses linguistic and visual information using a transformer while VAL [6] learns multiple transformers for the same at various levels through an attention mechanism. The objective function of VAL is designed to measure the feature similarities in a hierarchical manner. Hosseinzadeh *et al*. [20] compose images and text through locally bounded features (LBF). The state-of-the-art CosMo [25] models content and style changes between images and uses deep Multi-modal Non-Local (DMNL) [50] blocks to compose different types of changes.

Vision-Language Pretrained Transformers: Unlike transformers used in natural language processing [4, 12, 38], image classification [13, 30, 46], object detection [5, 58], and video understanding [14], Vision-Language Pre-trained (VLP) transformers are trained through self-supervision on large image-text corpora to capture prior multi-modal knowledge contained within them [8, 26–28, 44, 54, 57].

In this work, we apply VLPs to the problem of fashion image retrieval with textual feedback, so that we can benefit from the rich multi-modal information contained in their model weights. Our model is based on the state-of-the-art VLP VinVL [54], but is tailored with architectural additions for fashion retrieval and trained in a metric learning manner. Table 1 compares our model with previous works.

Table 1. Comparison of related works. The general VLP model VinVL is included for reference as our FashionVLP is based on this model. The columns *T*, *W*, *C*, *R*, and *L* refer to the inputs: text, whole image features, cropped clothing features, RoI features, and landmark features, respectively. AttNet refers to our new attention-based module for generating image encodings by fusing multiple contextual features.

	Reference					Target				
Method	T	W	C	R	L	W	C	L	**Reference Fusion**	**Target Feats.**
TIRG	✓	✓	✗	✗	✗	✓	✗	✗	Residual	CNN
VAL	✓	✓	✗	✗	✗	✓	✗	✗	Transformer	CNN
LBF	✓	✓	✗	✗	✗	✓	✗	✗	Cross-Attn	CNN
CosMo	✓	✓	✗	✗	✗	✓	✗	✗	DMNL	CNN
CIRPLANT	✓	✓	✗	✗	✗	✓	✗	✗	VLP	CNN
VinVL	✓	✓	✗	✓	✗	–	–	–	VLP	–
FashinVLP	✓	✓	✓	✓	✓	✓	✓	✓	VLP	AttNet

CIRPLANT

3. FashionVLP

As shown in Table 1, previous works in this domain are common in two aspects: (1) only the whole image is used as input, and is represented as global features after average pooling from convolutional feature-maps, and (2) customized modules are used for composing reference image and text features. However, both considerations are somewhat idealistic in the context of fashion image retrieval. More precisely, utilizing only whole fashion images implicitly requires robust feature extractors that generalize across fashion items with variations in size, rotation, pose, background, *etc*. Further, the use of global image features assumes that these are themselves sufficient and contain enough local information for retrieval. Using custom heuristic modules for image and text composition further raises concerns about generalization across fashion item types and variation in textual feedback. In order to ad-

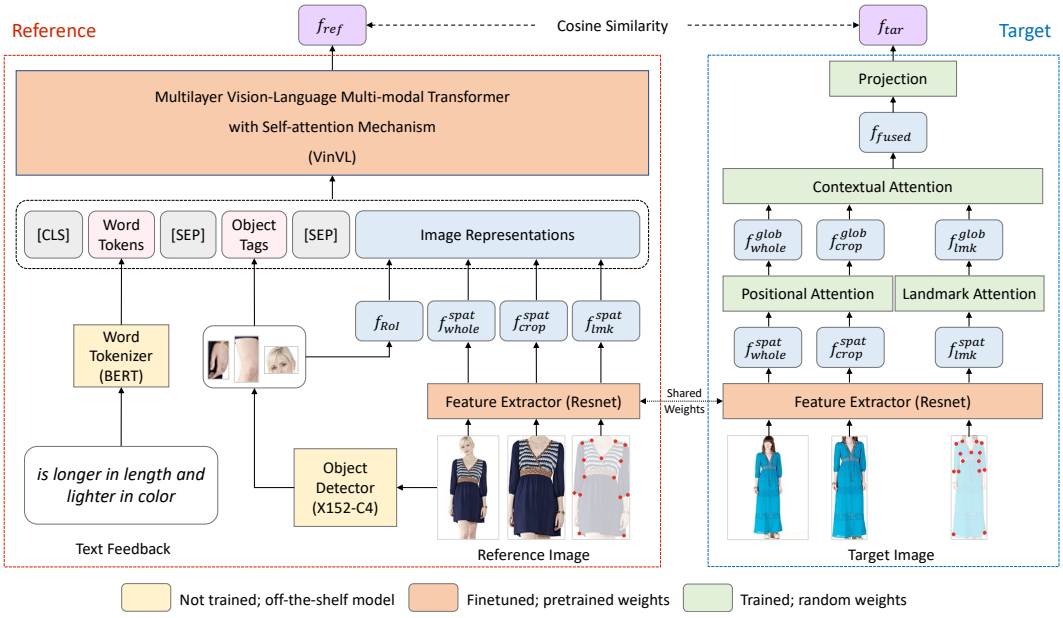


Figure 2. FashionVLP Overview. The model processes reference image-feedback pairs and target image candidates using parallel blocks – Reference and Target. Both blocks extract image features at multiple contextual levels, namely, whole image, cropped clothing, fashion landmarks, and regions of interest (RoIs), to focus on different fashion-related aspects of images. The Reference block fuses these image features with feedback inputs to generate joint reference embeddings f_{ref} through a transformer module that contains self-attention. The target block fuses image representations through multiple attention modules to generate target embeddings f_{tar}. The reference and target embeddings are then compared during training and inference for ranking candidate images for a query reference image and feedback pair.

dress both the considerations, we design a novel method for fashion retrieval with textual feedback.

Our method incorporates a VLP module for multi-modal information fusion as VLPs are known to generalize well across several domains [8, 26–28, 44, 54]. The inclusion of VLPs also brings the prior knowledge contained in large image-text corpora to the feedback-based fashion retrieval domain. Further, the core transformer design of VLPs allows for composition of additional modalities, *e.g.*, regions of interest (RoIs) [28, 54]. In order to better fit the fashion retrieval task, we introduce a novel overcomplete image representation that fuses multiple levels of fashion-related contextual information, namely, whole image, cropped clothing region, fashion landmarks, and regions of interest. Intuitively, such a representation provides direct and explicit inputs that are correlated with words in the feedback, and thus eases the fusion of linguistic and visual information, improving generalization.

3.1. Overview

Fashion image retrieval with textual feedback requires effective fusion of information between visual attributes of the reference image and linguistic content of the feedback. The tremendous success of Vision-Language Pre-trained (VLP) transformers at learning joint representations of such data makes them extremely suitable for this task. The efficacy of transformers is attributed to their self-attention mechanism, which allows information from non-adjacent inputs to be fused directly, unlike in traditional recurrent networks. Specifically, a transformer applies linear projections to its input features $X \in \mathbb{R}^{N \times d_{model}}$ to produce representations: $Q \in \mathbb{R}^{N \times d_k}$, $K \in \mathbb{R}^{N \times d_k}$ and $V \in \mathbb{R}^{N \times d_v}$. The output of a self-attention module is then computed as:

$$\text{Attn}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k})V \in \mathbb{R}^{N \times d_v} \quad (1)$$

The learning capacity of the attention module can be further improved by the multi-head design, formulated as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(h_1, h_2, \ldots, h_h)W^O, \quad (2)$$

$$h_i = \text{Attn}(Q_i, K_i, V_i). \quad (3)$$

where W^O is a linear layer that projects the concatenated features back to d_{model} dimensionality. The output of the attention module is post-processed [48] by a Feed-Forward Network and several Layer Normalization [2] layers.

Multi-layer transformers are built by stacking multiple transformer blocks sequentially. A typical example is BERT [12], which is designed for natural language tasks but has been extended to the multi-modal domain. Many VLP models are initialized with original BERT weights and further trained on domain-specific pre-training tasks [26, 28, 33].

In this work, we propose a new VLP model, FashionVLP, for fashion image retrieval with textual feedback. Figure 2 provides an overview of our framework. The basic

setup of the retrieval task requires learning representations of (a) the reference image and the textual feedback and (b) target images in a database in order to compare and search for candidate images to present to the user. Intuitively, our model consists of two parallel blocks – the reference block and the target block for encoding (a) and (b), respectively.

The reference block extracts features at multiple-levels of context – (1) whole image, (2) cropped image of clothing, (3) regions around fashion landmarks [32], and (4) regions of interest determined by a pretrained object detector. These features are then fused with object tags from the detector and word tokens from the feedback through a multi-layer transformer. The transformer output is treated as the final joint encoding f_{ref} for the reference image and feedback pair. The target block also computes encodings at contexts (1)–(3) using the same image feature extraction layers, but these are fused using a contextual attention mechanism instead of the transformer layers to reduce computation costs at inference, and projected to the dimensionality of f_{ref} to generate representations f_{tar} for target images. This allows for a scalable design for efficiently computing embeddings for fast-growing reference databases.

The model is trained using cosine similarity to compare reference and target embeddings. Subsequently, retrieval is performed by ranking candidate images using their similarities with the given reference image and textual feedback. In the following sections, we describe the computation of text and image features and the training methodology.

3.2. Linguistic Embedding

We tokenize the textual feedback through the pre-trained Oscar [28] tokenizer from VinVL [54]. The text is represented as a sequence of word tokens $t = \{w_1, w_2, \ldots, w_T\}$, where T is the length of the text. We append a special `[CLS]` token to the beginning of the sequence. When the feedback has mulitple sentences, we combine all the tokens into one sequence but separate sentences using `[SEP]` tokens. The tokens are then mapped to $\mathbb{R}^{T \times d_{model}}$ by an embedding layer. Finally, we add positional encoding to the sequence to preserve positional information.

3.3. Image Embeddings

Our model employs a ResNet [18] as the backbone feature extractor for whole, cropped, and landmark representations. We use a publicly available (https://git.io/JPAO4) pretrained Cascaded Pyramid Network [7] to extract fashion landmarks from input images. As shown in Figure 3, these landmarks are different for each clothing category and capture fashion-related semantics, *e.g.*, hem line, waist line, *etc*. Although not trained on shoes, the model is capable of effectively capturing meaningful points like tip, heel, *etc*.

Whole Image Representation: We obtain spatial image representations from the last convolutional block of a ResNet feature extractor with d_{img} channels by flattening it into a feature sequence $f^{spat}_{whole} \in \mathbb{R}^{HW \times d_{img}}$. The self-attention mechanism in the transformer layers of the reference block allows features from all positions on the feature map to be modeled simultaneously. Hence, we directly use this feature sequence as a part of the reference image representation. In the target block, however, we fuse these spatial features into global features. Global features are commonly computed by average-pooling spatial features, which fails to preserve location-specific salience. Therefore, we propose a positional attention module, a 1×1 convolution layer with d_{img} filters, to extract global representation $f^{glob}_{whole} \in \mathbb{R}^{d_{img}}$ from spatial features $f^{spat}_{whole} \in \mathbb{R}^{HW \times d_{img}}$.

$$f^{glob}_{whole} = \text{PositionalAttn}(f^{spat}_{whole}) * f^{spat}_{whole} \quad (4)$$

Cropped Clothing Representation: We use fashion landmarks to generate cropped clothing images from the given (whole) images in order to process their "zoomed in" versions through the feature extractor and better capture features from clothing regions. We then compute cropped clothing encodings in the same way as whole image embeddings. Specifically, the reference block computes f^{spat}_{crop} to provide as input to the transformer while the target block further generates f^{glob}_{crop} using positional attention.

Fashion Landmark Representation: We explicitly incorporate fashion semantics in our model by extracting feature maps corresponding to L fashion landmark positions from the second convolutional block of the ResNet feature extractor, which preserves more localized information. We then project these features to match the number of channels in the whole and cropped encodings, producing $f^{spat}_{lmk} \in \mathbb{R}^{L \times d_{img}}$. This is then directly used as input to the transformer as a part of the image representation in the reference block. However, in the target block, we use a landmark attention module, another 1×1 convolution with d_{img} filters, to combine $f^{spat}_{lmk} \in \mathbb{R}^{L \times d_{img}}$ and generate $f^{glob}_{lmk} \in \mathbb{R}^{d_{img}}$

RoI-level Representation: Due to the size of image-text corpora typically used to train VLPs, CNNs are usually not integrated into the framework. Instead, existing models ex-

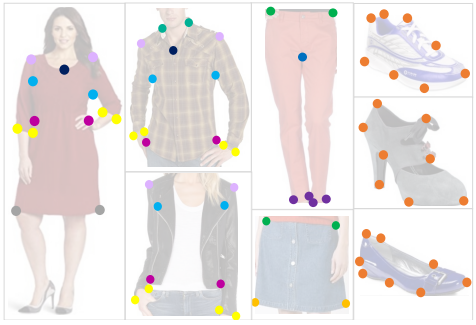


Figure 3. Fashion landmarks visualization for different clothing types. Landmarks reflect essential points such as neckline, armpits, *etc*., that provide useful visual cues for fashion retrieval.

tract RoI-level features through a pre-trained object detector. The semantic information in RoIs is indeed crucial for feedback-based fashion image retrieval. For instance, when a customer asks for changes on sleeves, a model with RoI information would be able to place higher attention on RoIs corresponding to arms. Therefore, we include RoI-level features, f_{RoI}, extracted by a pre-trained object detector as part of our image representation in the reference block.

We use a publicly available [54] Faster-RCNN-C4 [39] with ResNeXt-152 [52] backbone (X152-C4) trained on MSCOCO [29], Visual Genome [23], Objects365 [43], and OpenImage V5 [24], following [54]. The RoIs are filtered by a confidence threshold ϵ. In addition, [28, 54] state that the object category for each region can acts as an anchor between images and text. We follow the setting in [28] and append object tags to the end of the linguistic input in the reference block, separated by the `[SEP]` token.

Region Position Encoding: In order to preserve positional information of the extracted RoI features, we encode their region position into a 6-dimensional vectors as

$$f_{pos} = \left[\frac{x_1}{w}, \frac{y_1}{h}, \frac{x_2}{w}, \frac{y_2}{h}, \frac{x_2 - x_1}{w}, \frac{y_2 - y_1}{h}\right], \quad (5)$$

where $[x_1, y_1, x_2, y_2]$ denotes the bounding box of the RoI and h, w are image dimensions. We combine f_{pos} with f_{RoI} to create position-aware region representations. We similarly combine f^{spat}_{whole} and f^{spat}_{crop} with their corresponding position encodings according to their field of view.

Combined Reference Image Representation: The final image representation fed into the transformer layers in the reference block consists of (1) f^{spat}_{whole}, (2) f^{spat}_{crop}, (3) f_{ROI}, (4) f^{spat}_{lmk}, and (5) position encodings for (1)–(3).

Fused Target Representation: We combine f^{glob}_{whole}, f^{glob}_{crop}, and f^{glob}_{lmk} using a contextual attention module, which is a 1×1 convolution layer, to get a fused target representation.

3.4. Model Training

The transformer takes three input segments: linguistic features, object tags, and image features. We take the output hidden state of the `[CLS]` token from the transformer as the reference embedding f_{ref}. Meanwhile, for target representation we extract the fused features and project it into the joint feature space to get f_{tar}. We then compute the similarity of f_{ref} and f_{tar} by a kernel function κ.

We adopt a batch-based classification loss [49], where each entry inside a batch acts as a negative sample for all other entries. This objective function converges faster [42] than triplet loss, especially on complex datasets. For a batch of B image-text pairs, the loss is defined as:

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp(\kappa(f^i_{mod}, f^i_{tar}))}{\sum_{j=1}^{B}\exp(\kappa(f^i_{mod}, f^j_{tar}))}. \quad (6)$$

The kernel κ in Equation (6) can be any metric, but we use inner product in this work, resulting in cosine similarity.

We train our network by fine-tuning the transformer together with the feature extractor and the attention modules. The feature extractors in reference and target blocks share weights to prevent overfitting. We do not fine-tune the object detector as the Region Proposal Network [39] inside X152-C4 cannot be trained without separate loss functions.

4. Evaluation

We evaluate models on FashionIQ [51], Shoes [3], and Fashion200K [17]. We compare our model with state-of-the-art methods: TIRG [49], VAL [6], and CosMo [25]. We additionally present results of visual reasoning based baselines: RN [40], MRN [21], and FiLM [36]. In the following sections, we describe the experiment setup, present evaluation results, and discuss ablation studies.

4.1. Experiment Setup

Implementation Details: We use ImageNet [11] pretrained ResNet-50 for FashionIQ and Shoes, and ResNet-18 for Fashion200K, as image feature extractors following [25]. We employ BERT-base [12] from [54] as the transformer. We use the Adam [22] optimizer with $\beta = (0.55, 0.999)$ and train models for 100 epochs, halving the learning rate every 10 epochs. We use batch sizes of 80 and 92 for FashionIQ and Shoes, respectively, with an initial learning rate of $4\text{e}{-4}$ and a warm-up period of 150 iterations. We set batch size as 200, initial learning rate as $1\text{e}{-3}$, and warm-up period as 500 iterations for Fashion200K due to its large size. The detection confidence threshold ϵ is set to 0.5.

Inference: During inference, we process queries and candidates in the dataset separately. Candidate features are extracted by the target block containing only image feature extractors and attention modules, while reference and feedback queries are processed by the reference block including the transformer module as described in Section 3.1. We then compute cosine similarities for ranking the candidates.

Evaluation Metric: Models are evaluated using the standard top-K recall metric for image retrieval, denoted as R@K. Performance is compared specifically on the average of R@10 and R@50 as a metric of overall performance.

4.2. Results

FashionIQ [51]: This is a fashion retrieval dataset with interactive natural language captions. Items belong to three types: Dresses, Tops&Tees, and Shirts. It contains 77K images in total with 46K images for training and 18K image pairs available. Each pair has two crowdsourced captions that describe changes from the reference to the target. The feedback is complicated and sometimes a sentence includes multiple concepts to be changed, *e.g.*, "is patterned and has

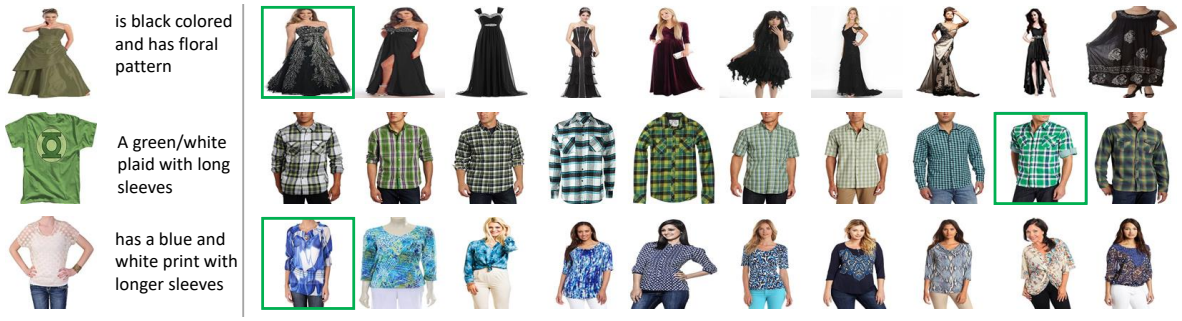


Figure 4. Qualitative results on FashionIQ. We show reference images on the left and top-10 retrievals with descending scores on the right. Ground-truths are shown with boxes. Feedback in FashionIQ is complex yet realistic and can contain multiple concepts simultaneously.

Table 2. Quantitative results on FashionIQ. Our model surpasses the state-of-the-art by a large margin on all three sub-categories. We report results with both the VAL evaluation protocol [6,25] and the Original evaluation protocol. CT denotes CIRPLANT [31].

	Dress		Toptee		Shirt		Overall		
Method	**R@10**	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	**Mean**
VAL [6, 25] Evaluation Protocol									
RN [40]	15.44	38.08	21.10	44.77	18.33	38.63	18.29	40.49	29.39
MRN [21]	12.32	32.18	18.11	36.33	15.88	34.33	15.44	34.28	24.86
FiLM [36]	14.23	33.34	17.30	37.68	15.04	34.09	15.52	35.04	25.28
TIRG [49]	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	27.40
CT [31]	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
VAL [6]	21.12	42.19	25.64	49.49	21.03	43.44	22.60	45.04	33.82
CosMo [25]	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.31	39.45
FashionVLP	**32.42**	**60.29**	**38.51**	**68.79**	**31.89**	**58.44**	**34.27**	**62.51**	**48.39**
Original Evaluation Protocol									
Image Only	4.46	13.19	5.46	13.21	6.13	13.64	5.35	13.35	9.35
Concat	14.92	34.95	14.28	34.73	12.71	30.08	13.92	33.25	23.59
TIRG [49]	14.13	34.61	14.79	34.37	13.10	30.91	14.01	33.30	23.66
CosMo [25]	21.39	44.45	21.32	46.02	16.90	37.49	19.87	42.62	31.25
FashionVLP	**26.77**	**53.20**	**28.51**	**57.47**	**22.67**	**46.22**	**25.98**	**52.30**	**39.14**

a halter neckline", "is black with floral patterns", *etc*. The complex yet realistic nature of the feedback in this dataset makes it exceptionally challenging for the retrieval task.

We follow the evaluation protocol of [6, 25], where the candidate set is constructed by unifying all reference and target images in the test set. This reduces the number of images for retrieval, compared with the *original* test set, resulting in higher performance for all models. We evaluate models on the reduced set (VAL [6] evaluation protocol) for fair comparison with previous works, but also report results for the *original* evaluation protocol for future reference.

Quantitative results are presented in Table 2. Our model outperforms the previous state-of-the-art by a large margin on all metrics. Specifically, for the VAL evaluation protocol, our approach achieves a relative improvement of more than 29% on R@10 and 19% on R@50. Furthermore, we observe a broad 23% relative improvement over all fashion types, indicating that our model generalizes well across them. Finally, FashionVLP also shows an overall 25% relative improvement for the *original* evaluation protocol.

Figure 4 presents some examples of retrieval. As shown, feedback sentences in FashionIQ are complex and contain multiple concepts. Our model is able to capture such diverse concepts and retrieve good candidate images.

Table 3. Quantitative results on Fashion200K. Our model achieves the best results on Recall@50 and mean recall.

Method	**R@10**	**R@50**	**Mean**
RN [40]	40.5	62.4	51.4
MRN [21]	40.0	61.9	50.9
FiLM [36]	39.5	61.9	50.7
TIRG [49]	42.5	63.8	53.2
VAL [6]	49.0	68.8	58.9
CosMo [25]	**50.4**	69.3	59.8
FashionVLP	49.9	**70.5**	**60.2**

Fashion200K [17]: This is a large-scale fashion dataset with images from various online shopping websites. It contains more than 200K images (training: 172K, testing: 33K) and a feedback vocabulary of more than 5K words. Images are labeled with descriptions like "blue women's embroidered midi-dress", and attributes including product information and user reviews. In our experiments, we only utilize images and their descriptions. Following [49], we generate textual feedback through an automated process that compares attributes between pairs of images. The feedback is structured in the form of "*replace [sth] with [sth]*", which is *much simpler* than feedback in FashionIQ and Shoes.

Our results presented in Table 3 show that our model outperforms the previous state-of-the-art by a relative improvement of 1.7% on R@50. Although R@10 is slightly lower, our model achieves an overall relative improvement of 0.6%. We attribute smaller gains on Fashion200K to the *fixed unnatural templated nature* of feedback in this dataset, as shown in Figure 5. Such text is closer to attribute-like feedback [6] than to natural language sentences. As FashionVLP aims to bring the benefits of strong natural language priors to the task of fashion retrieval, such knowledge is not as beneficial for Fashion200K. However, our model still achieves the best results on this dataset.

Qualitative results in Figure 5 show that multiple images are considered correct for a query if their captions are identical. Results show that our model can recognize attribute changes in the feedback and retrieve images accordingly.

Figure 5. Qualitative results on Fashion200K. We show reference images on the left and top-10 retrievals with descending scores on the right. Ground-truths are shown with boxes. Note that a query pair can correspond to multiple valid target images in this dataset. Due to the lack of human annotated feedback, comments in Fashion200K follow the template: *replace [sth] with [sth]*, and are thus less instructive.

Figure 6. Qualitative results on Shoes. We show reference images on the left and top-10 retrievals with descending scores on the right. Ground-truths are shown with boxes. Feedbacks in Shoes are fine-grained and contain concepts belonging to the fashion domain of shoes.

Table 4. Quantitative results on Shoes. Our model achieves the best results on Recall@50 and mean recall.

Method	R@10	R@50	Mean
RN [40]	45.10	71.45	58.27
MRN [21]	41.70	67.01	54.35
FiLM [36]	38.89	68.30	53.59
TIRG [49]	45.45	69.39	57.32
VAL [6]	**49.12**	73.53	61.32
CosMo [25]	48.36	75.64	62.00
FashionVLP	49.08	**77.32**	**63.20**

Shoes [3]: This dataset was originally collected to extract attribute information from web images. Guo *et al*. [16] tagged the images with captions in natural language for fashion image retrieval. We use the original splits in [16], which provides 10K training pairs and 4.6K test queries.

Table 4 shows that our model achieves the best results on this dataset, with relative improvements of 2.2% on R@50, 1.5% on R@10, and 1.9% on average. Qualitative results in Figure 6 show that our model can perceive both simple visual changes like color and complex visual properties like patterns and shoe models for retrieving candidate images.

4.3. Ablation Studies

We present ablation studies to provide insights into how different contextual image information and model components affect performance. We perform these studies on FashionIQ as it contains complex and realistic feedback.

Table 5. Ablation study on FashionIQ on different contextual image features. *PositionalAttn*, *RoI*, *Lmk*, *Crop* and *Whole* refer to positional attention, RoI encodings, landmark features, and embeddings from cropped and whole images, respectively.

Method	R@10	R@50	Mean
FashionVLP	**34.27**	**62.51**	**48.39**
w/o PositionalAttn	33.75	61.43	47.59
w/o Lmk	33.28	60.77	47.02
w/o Lmk, Crop	32.60	59.75	46.18
w/o Lmk, Crop, RoI	31.67	60.06	45.86
w/o Lmk, Crop, Whole	31.34	59.84	45.59

Contextual image features: We analyze the effect of positional attention, landmark, cropped clothing, and RoI features on the retrieval performance by evaluating versions of our model trained without these encodings. Results in Table 5 show that excluding these contextual pieces reduces performance. Specifically, using global average pooling instead of positional attention to combine spatial features results in a 1.7% relative reduction in mean recall. Removing landmark features causes a 2.83% relative drop. Additionally excluding cropped clothing encodings results in a 4.57% drop. Further removing RoI features causes 5.23% degradation. Excluding whole image encodings instead of RoI features as in VinVL [54] leads to a 5.79% relative drop.

Fashion landmark features and fusion methods: We first study the effects of different methods of generating landmark features: (1) normalized landmark coordinates, (2)

Figure 7. Visualization of attention on relevant words in textual feedback and different contextual image features for two sample pairs. Words with the highest attention weights are shown in bold. For each level of context in the reference block, we visualize the attention heatmap of the corresponding most attended word, and observe effective correspondence between bold words and relevant image regions. On the target side, we visualize attention heatmaps corresponding to our positional and landmark attention modules, showing that these modules effectively capture important fashion information. Results of attention in the reference (left) and the target (right) blocks further show that the whole image modality is insufficient – for example, the upper sample's whole image representation for the target image lacks any useful information. Further, fashion landmarks provide important for fashion-specific concepts, *e.g.*, *strap*, *hem*, *etc*.

Table 6. Ablation study on FashionIQ on different methods of generating landmark representations (LmkRep) and combining them. *Conv Block2* and *Conv Block3* indicate that features for each landmark are extracted from the 2nd and the 3rd convolutional blocks of the feature extractor, respectively. *Norm coords* refers to the use of normalized landmark positions as feature values. For fusion, we compare the effects of the context (Ctx) and the landmark (Lmk) attention (Attn) modules with simply concatenating the features.

LmkRep	Ctx Attn	Lmk Attn	R@10	R@50	Mean
Conv Block2	✓	✓	**34.27**	**62.51**	**48.39**
	✓	✗	33.17	61.42	47.29
	✗	✗	32.15	61.09	46.62
Conv Block3	✓	✓	33.63	61.85	47.74
	✓	✗	32.09	60.48	46.29
	✗	✗	33.81	61.12	47.46
Norm coords	✓	–	32.82	61.02	46.92
	✗	–	32.70	61.10	46.90

indexed features from the third convolutional block of the ResNet feature extractor, and (3) those from the second convolutional block. Results in Table 6 show that using features from the lower (second) block achieves the best performance, indicating that the fine-grained local information provided by this block is useful for fashion retrieval.

We also study the effects of different methods of fusing image features in the target block. Adding contextual and landmark attention to combine whole, cropped clothing, and landmark features (second convolutional block) provides 3.8% relative improvement compared to simply concatenating the said features. Of this, incorporating the landmark attention module for fusing landmark features before the contextual attention provides 2.3% improvement.

Attention Visualization: In order to further analyze the above two ablation studies, we visualize attention maps in Figure 7. For reference images, we first extract the most attended words from query text and then visualize their corresponding attention on image features. We find that RoIs features best capture broad concepts,*e.g.*, *design* and *dress*, whereas fashion landmarks are useful for specific attributes,*e.g.*, *hem* and *straps*. Cropped clothing features provide access to zoomed-in regions like *neck*. Our model is also able to reason about ambiguous concepts like *spaghetti* and focus on relevant parts of image. On target side, adding spatial attention helps remove irrelevant information and focus on important regions like cloth design and sleeves.

5. Conclusion

We have presented a new vision-language transformer based model, FashionVLP, which leverages prior knowledge from large image-text corpora and multiple contextual image features to effectively perform fashion image retrieval with textual feedback. Our model also provides a novel attention based approach for effectively fusing visual information from diverse visual contexts for learning candidate image embeddings. Furthermore, we present an efficient framework for generating embeddings of candidate images, which are in the same latent space as the joint encodings of reference image and feedback queries, by excluding the parameter-heavy transformer layers from the computation process. Results show that our model achieves state-of-the-art results on benchmark datasets. In particular, our model achieves more than 23% relative improvement on FashionIQ, which contains complex yet realistic natural language feedback for fashion image retrieval.

Despite our novel design, there are still challenges in implementing it for real-life fashion search applications involving in-the-wild user queries, *e.g.*, a fashion item could be hung in a closet, folded in a box, *etc*., with diverse conditions, *e.g.*, old, wet, *etc*. Moreover, common challenges like lighting, color distortion, motion blur, *etc*., still exist. Finally, the current formulation is limited to a single round of retrieval based on one reference image and feedback pair, while real-life use-cases would require multiple iterations of feedback involving some form of memory management.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[3] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010. 2, 5, 7

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[6] 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)*, June 2020. 1, 2, 5, 6, 7

[7] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 4

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 1, 2, 3

[9] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 2

[10] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[12] 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. 2, 3, 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 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*, pages 241–257. Springer, 2016. 2

[16] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 7

[17] 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 international conference on computer vision*, pages 1463–1471, 2017. 2, 5, 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[19] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[20] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3596–3605, 2020. 2

[21] 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*, pages 361–369, 2016. 2, 5, 6, 7

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

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 5

[24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 5

[25] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feed-

back. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1, 2, 5, 6, 7

[26] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 1, 2, 3

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2, 3

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2, 3, 4, 5

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[30] 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. *arXiv preprint arXiv:2103.14030*, 2021. 2

[31] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, October 2021. 2, 6

[32] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision*, pages 229–245. Springer, 2016. 1, 4

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 1, 3

[34] Andres Mafla, Rafael S. Rezende, Lluis Gomez, Diane Larlus, and Dimosthenis Karatzas. Stacmr: Scene-text aware cross-modal retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2220–2230, January 2021. 2

[35] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 2

[36] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2, 5, 6, 7

[37] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016. 2

[38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 5

[40] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 5, 6, 7

[41] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5814–5824, 2019. 2

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 5

[43] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019. 5

[44] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1, 2, 3

[45] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2

[47] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[49] 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*, pages 6439–6448, 2019. 1, 2, 5, 6, 7

[50] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[51] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 11307–11317, 2021. 2, 5

[52] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[53] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12):5412–5425, 2020. 2

[54] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, June 2021. 1, 2, 3, 4, 5, 7

[55] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3536–3545, 2020. 2

[56] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020. 2

[57] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 1, 2

[58] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

[59] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. *CoRR*, abs/2103.16110, 2021.