Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models

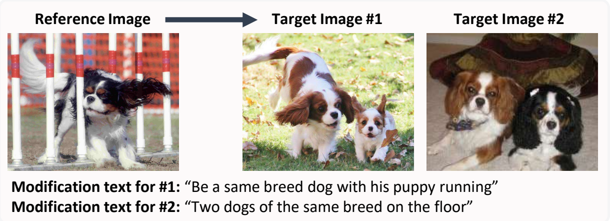
Zheyuan Liu[1] Cristian Rodriguez-Opazo[2] Damien Teney[2,3] Stephen Gould[1]
[1]Australian National University
[2]Australian Institute for Machine Learning, University of Adelaide [3]Idiap Research Institute
{zheyuan.liu, stephen.gould}@anu.edu.au
cristian.rodriguezopazo@adelaide.edu.au, damien.teney@idiap.ch

Abstract

We extend the task of composed image retrieval, where an input query consists of an image and short textual description of how to modify the image. Existing methods have only been applied to non-complex images within narrow domains, such as fashion products, thereby limiting the scope of study on in-depth visual reasoning in rich image and language contexts. To address this issue, we collect the Compose Image Retrieval on Real-life images (CIRR) dataset, which consists of over 36,000 pairs of crowd-sourced, open-domain images with human-generated modifying text. To extend current methods to the open-domain, we propose CIRPLANT, a transformer based model that leverages rich pre-trained vision-and-language (V&L) knowledge for modifying visual features conditioned on natural language. Retrieval is then done by nearest neighbor lookup on the modified features. We demonstrate that with a relatively simple architecture, CIRPLANT outperforms existing methods on open-domain images, while matching state-of-the-art accuracy on the existing narrow datasets, such as fashion. Together with the release of CIRR, we believe this work will inspire further research on composed image retrieval. Our dataset, code and pre-trained models are available at https://cuberick-orion.github.io/CIRR/.

1. Introduction

We study the task of *composed image retrieval*, that is, finding an image from a large corpus that best matches a user query provided as an image-language pair. Unlike traditional content-based [38] or text-based [24, 42] image retrieval where a single modality is used to describe the target image, composed image retrieval involves both visual and textual modalities to specify the user's intent. For humans the advantage of a bi-modal query is clear: some concepts and attributes are more succinctly described visually, others through language. By cross-referencing the two modalities, a reference image can capture the general gist of a scene, while the text can specify finer details. The challenge is the inherent ambiguity in knowing what information is important (typically one object of interest in the scene) and what can be ignored (*e.g*., the background and other irrelevant objects). However, existing datasets for this task fall short of allowing us to adequately study this problem.



Modification text for #1: "Be a same breed dog with his puppy running"
Modification text for #2: "Two dogs of the same breed on the floor"

Figure 1. Example of composed image retrieval from the proposed CIRR dataset. The input is composed of a reference image and a modifying text, to which the model must find a close match. A major challenge is the inherent ambiguity and underspecification of visual aspects to be preserved or modified. Our dataset includes open-domain images with rich contexts to facilitate the study of such challenge.

Consider the example in Fig. 1. Real-life images usually contain rich object interactions on various scales. In each case, to readily identify the relevant aspects to keep or change and pay less attention elsewhere (*e.g*., the color of the dog's fur and background objects), a model must develop in-depth visual reasoning ability and infer implicit human agreements within both the visual and language contexts. However, existing datasets are constrained to domains such as fashion products [4, 12, 13] or synthetic objects [40] with relatively simple image contents. We argue that the current datasets are insufficient for exploring the unique research opportunity mentioned above.

Motivated by this problem, we collect the Compose Image Retrieval on Real-life images (CIRR) dataset. It is based on the open-domain collection of real images from NLVR^2 [35], for which we collected rich, high-quality annotations that aim to tease out the important aspects of the reference image and textual description for a given query.

Compared with existing datasets, CIRR places more emphasis on distinguishing between visually similar images, which provides a greater challenge, as well as a chance for studying fine-grained vision-and-language (V&L) reasoning in composed image retrieval. Our dataset also allows for evaluation on fully labeled subsets, which addresses a shortcoming of existing datasets that are not fully labeled and therefore contain multiple false-negatives (as unlabeled images are considered negative).

Meanwhile, we propose Composed Image Retrieval using Pretrained LANguage Transformers (CIRPLANT), which extends current methods into open-domain images by leveraging the knowledge of large-scale V&L pre-trained (VLP) model [25]. Although the advantages of such pretrained models have been validated in many visiolinguistic tasks [6, 25, 28], to the best of our knowledge, none have been applied to composed image retrieval. We conjecture one of the reasons being the existing domain-specific datasets cannot greatly benefit from the pre-training, which uses more complex, open-world images. Moreover, to adopt the VLP models for fine-tuning, most of the downstream tasks are formulated as classification tasks [6, 25]. For composed image retrieval, it requires taking as input both the reference and target images. However, this greatly raises the computational overhead for retrieval, as the model needs to exhaustively assess each input query paired with each candidate target before yielding the one with the highest prediction score. Instead, we propose to preserve the conventional metric learning pipeline, where the input queries are jointly embedded using the VLP model and later compared with features of candidate images through ℓ_2-norm distance. Specifically, our design maintains the same objective of "language-conditioned image feature modification" as previous work [5, 8, 40], while manages to utilize the pre-trained V&L knowledge in large-scale models. We demonstrate that our proposed model reaches state-of-the-art on the existing fashion dataset while outperforming current methods on CIRR.

2. Related Work

Image retrieval. Existing work on image retrieval using deep learning can be categorized by the type of queries considered. Content-based Image Retrieval (CBIR) refers to the use of image-only queries for product search [26], face recognition [29, 34], etc. This setup leaves little room for iterative user feedback or refinement. Other possible modalities to form queries include attributes [13], natural language [24, 42], and sketches [31]. These are motivated by a more natural user experience, but require more advanced retrieval mechanisms. Vo et al. [40] propose *composed image retrieval* that combines visual and text modalities. Here the query consists of a reference image and short text describing desired differences with this image. Guo et al. [12] demonstrate the potential of this setup for the narrow domain of fashion recommendation.

Our work focuses on composed image retrieval in an open-domain setting, *i.e*., not restricted to fashion products for example. We specifically address the case of distinguishing visually similar images, which requires more in-depth, fine-grained reasoning ablility over both the visual and language modalities.

Compositional learning. The topic of compositional learning has been extensively studied in V&L tasks including visual question answering (VQA) [3], image captioning [1, 2] and video retrieval [41]. The aim is to produce learned joint-embedding features that capture the salient information in both visual and text modalities along with their interactions. For composed image retrieval, Vo et al. [40] first propose a residual-gating mechanism that aims to control variation of the input image features through text. Hosseinzadeh and Wang [17] use region-based visual features from R-CNN models [10, 32] originally proposed for image captioning [1] and VQA [37]. Recently, Chen et al. [5] use a transformer-based model [39] and inject the text modality at varying depths of the image model. Dodds et al. [8] introduce the concept of modality-agnostic tokens, which they obtain from "divided" spatial convolutional features and LSTM hidden states. In this work, we propose a method that leverages the rich knowledge in VLP models. Our method can modify the input image features based on natural language without the need of developing monolithic architecture on the specific task.

Vision-and-language pre-training. The success of pretrained BERT [7] inspired numerous attempts on VLP models, including [6, 23, 25, 28, 36]. The aim is to develop Transformer-based [39] models trained on large-scale image-text triplets to produces V&L representations applicable to various tasks. The advantage is clear, instead of training monolithic models on task-specific datasets from zero, different V&L tasks can start with the representations learned from (usually) a considerably larger image-text corpus, and fine-tune on specific tasks. Motivated by success in other V&L tasks, we propose to adopt the VLP model on composed image retrieval. The key obstacle is to design the architecture to encourage a controlled modification of image features, which, differs greatly from the conventional use cases of such models.

Datasets for composed image retrieval. Most existing datasets suitable for composed image retrieval are repurposed from other tasks [13, 18, 40]. Images are paired within classes and textual descriptions of their differences are generated automatically from existing labels. These datasets are relatively simple visually and only contain short descriptions with simple language. CSS [40] uses the synthetic images of geometric 3D shapes from CLEVR [20], paired with descriptions generated according to differences in appearance of the objects. Fashion200k [13] contains approx. 200k images tagged with attributes that can be used to compose text descriptions of differences between images. MIT-States [18] contains images of entities in different states each labelled with one noun and one adjective. The adjectives can describe limited differences between images. More recent works introduced human-generated descriptions. Guo et al. [11] present annotations for Shoes [4], a dataset of 10k footwear images. Fashion-IQ [12] contains crowd-sourced descriptions of differences between images of fashion products. Dodds et al. [8] introduce benchmarks for the Birds-to-Words [9] and Spot-the-Diff [19] datasets.

In this paper, we introduce a new dataset that addresses current deficiencies. Our dataset is open-domain and not restricted, *e.g*., to fashion products [4, 12, 13]. We design a careful collection process to produce high-quality pairs from our diverse collection of images by only associating visually- and semantically-related images. We also address the issue of false-negative targets, that is, candidate target images that are valid for a certain input query, but not labeled as such. Previous datasets failed to resolve this issue due to the cost of exhaustively labeling images against every possible query, which is mitigated by our data collection strategy. Although not used in our current work, the dataset also contains a rich set of auxiliary annotations that clarify ambiguities not addressed in the textual query.

3. The Proposed Model

In this section, we first briefly introduce the vision-and-language pre-trained (VLP) models, then we discuss our adaptation of it for the task of composed image retrieval.

3.1. Vision-and-Language Pre-trained Models

Contemporary VLP models are inspired by BERT [7], which is constructed with multi-layer transformers [39]. The model accepts variable-length sequential inputs $\boldsymbol{i}_{\text{VLP}}$, which consist of a concatenation among words in the text sequence(s) $\boldsymbol{w} = \{w_1, \ldots, w_T\}$, regional features from the image $\boldsymbol{v} = \{v_1, \ldots, v_K\}$, and other optional tokens. For instance, in OSCAR [25], an object label associated with each regional feature is appended to the end as $\boldsymbol{l} = \{l_1, \ldots, l_K\}$.

Within each transformer layer, a multi-head self-attention mechanism is designed to capture the dependencies among the sequential tokens. Layers are stacked hierarchically to attend to the output of the previous layer. Once pre-trained on a large corpus, the final output representations can be used for fine-tuning on arbitrary downstream tasks, where the usage varies depending on the task.

That said, downstream tasks share some common aspects. Mostly, a classification token `[CLS]` is inserted at the start of the input text sequence, which aggregates information from the modalities. The final `[CLS]` output is then used to make predictions, such as for image classification.

3.2. Adaptation to Composed Image Retrieval

The task of composed image retrieval can be formally described as finding the target image in a large corpus of images $I_{\text{T}} \in \mathcal{D}$ that best matches a query provided by a reference image-text pair $q = \langle I_{\text{R}}, t\rangle$. Our goal is to learn a text-image composition module, which maps a given $\langle I_{\text{R}}, t\rangle$ into the same embedding space as, and close to, the corresponding I_{T}. Intuitively speaking, this requires the composition module to modify I_{R} conditioned on t.

In this work, we employ OSCAR [25], a recently proposed VLP model with state-of-the-art performance as the composition module to perform the mapping as follows.

Input sequence. We denote the input sequence of OSCAR as $\boldsymbol{i}_{\text{VLP}} = \{\boldsymbol{w}, \boldsymbol{v}\}$, where we initialize OSCAR without the optional object label inputs $\boldsymbol{l}$. We then follow Li *et al.* [25] for processing text sequences, but introduce the following adaptations on image representations.

Rather than including a set of regional features, we pre-process images through an ImageNet pre-trained ResNet [14] model and extract features from before the final FC-layer. We then process these features through a (newly) learned FC-layer and ℓ_2-normalization to give a single image feature $\boldsymbol{v} = \{v_1\}$ as the input to OSCAR. This same feature representation is used for the corpus of candidate target images $I'_{\text{T}} \in \mathcal{D}$ as shown in Fig. 2.

We choose this relatively simple design for two reasons. First, recent work (*e.g*., [16]) has shown the compatibility between VLP models and non-regional features of images. Second, we hypothesize that using global image features is easier to achieve our goal of modifying I_{R} conditioned on t so as to closely match I_{T}.

Output token. As shown in Fig. 2, contrary to typical downstream tasks, we do not use the final representation of the `[CLS]` token as the text-image joint embedding. Instead, we extract the representation corresponding to the image feature token and treat it as the composed image-text feature. This resembles the fine-tuning of REF [23], as well as VLN-BERT [16]. In both cases, tokens other than `[CLS]` are used for prediction. For composed image retrieval, our design makes sense since the transformer model includes residual connections between input and output tokens. Intuitively, the reference image features are *modified* by aggregating the information from other word tokens to produce the target image features.

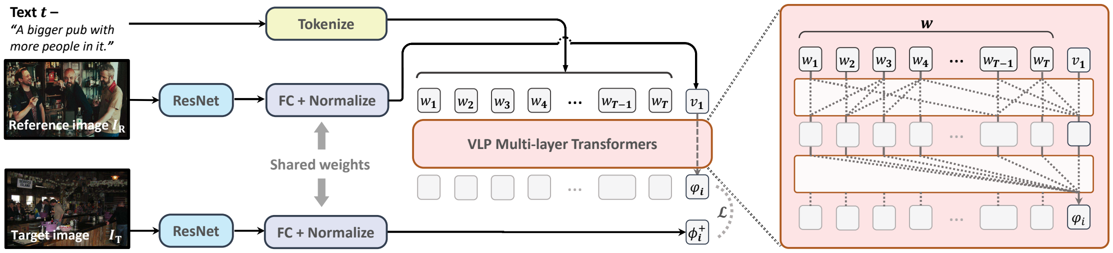


Figure 2. (Left) Schematic of our model. Given a pair of reference image and text as input, we aim at learning a *modified* image feature of the reference image conditioned on the text, such that it matches the feature of the target image. To compare image features of reference and candidate target images, we extract ResNet features and use a shared FC-layer (with normalization) to project them into the same domain. (Right) Overview of the image-text composition module using vision-and-language pre-trained (VLP) multi-layer transformers. Dashed lines (not fully drawn) represent feature aggregation by attention, which learns a language-conditioned image feature modification.

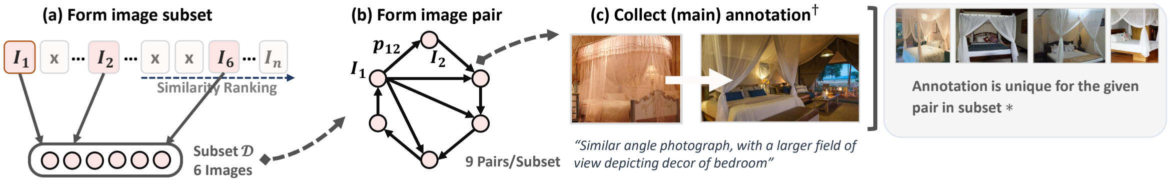

(a) Form image subset
I_1 x … I_2 … x x I_6 … I_n
Similarity Ranking
Subset $\mathcal{D}$
6 Images
(b) Form image pair
p_{12}
I_1
I_2
9 Pairs/Subset
(c) Collect (main) annotation†
"Similar angle photograph, with a larger field of view depicting decor of bedroom"
Annotation is unique for the given pair in subset ∗

Figure 3. Overview of the data collection process. (a) We demonstrate the construction of an image subset. (b) We illustrate how we choose and form 9 image pairs within one subset, where each arrow suggests the direction from a reference to a target image. (c) † represents Human Tasks with AMT workers. ∗ indicates the instruction that mitigates the issue of false-negative.

Metric learning. We use soft triplet-based loss with ℓ_2-norm distance as in Vo *et al.* [40] to bring the composed image-text feature closer to the feature of the target image (positive pair), while pulling apart the features of negative pairs. In essence, given the i-th positive pair $\langle \varphi_i, \phi_i^+ \rangle$ and an arbitrary negative $\phi_{i,j}^-$ among all negatives ϕ_i^-, the loss is computed as:

$$\mathcal{L} = \log[1 + \exp(\kappa(\varphi_i, \phi_{i,j}^-) - \kappa(\varphi_i, \phi_i^+))], \qquad (1)$$

where κ is ℓ_2-norm distance. In training, we randomly sample the negative for each pair and average the loss over all sampled triplets $\langle \varphi_i, \phi_i^+, \phi_{i,j}^- \rangle$.

4. The CIRR Dataset

Existing datasets for composed image retrieval [12, 40] contain training and testing examples as triplets $\langle I_\text{R}, q, I_\text{T} \rangle$ where $q = \langle I_\text{R}, t \rangle$ forms the query and I_T is (an example of) the desired target from a large image corpus $\mathcal{D}$. However, these existing datasets have two major shortcomings. First, they lack the sufficient visual complexity to facilitate the study of one of the major challenges in composed image retrieval, which is the subtle reasoning over what aspects are important and what shall be ignored. Second, since the candidate images cannot be extensively labeled for each $\langle I_\text{R}, t \rangle$ pair, existing datasets contain many false-negatives. That is, images $I \in \mathcal{D}$ that are valid matches for the query but not labeled as the ground-truth target I_T. Indeed, all images in $\mathcal{D} \setminus \{I_\text{R}, I_\text{T}\}$ are considered as negatives. To circumvent this shortcoming, existing works choose to evaluate models with Recall@K and set K to larger values (*e.g.*, 10, 50 [12]), thus accounting for the presence of false-negatives. However, the issue persists during training. Moreover, by setting larger K values, these methods are essentially trading in their ability for learning detailed text-image modifications.

To mitigate these issues, we introduce the Compose Image Retrieval on Real-life images (CIRR) dataset, which includes over 36,000 annotated query-target pairs, $\langle q = \langle I_\text{R}, t \rangle, I_\text{T} \rangle$. Unlike existing datasets, we collect the modifying text to distinguish the target from a set of similar images (addressing the problem of false-negatives) and creating challenging examples that require careful consideration of visual and textual cues. Details are as follows.

4.1. Data Collection

We first form image pairs then collect related annotations by crowd-sourcing. The pairs are drawn from subsets of images, as described below. This strategy plays a major role in mitigating the issue of false negatives (see Sec. 5). Fig. 3 outlines our data collection procedure.

Image source. We use the popular NLVR^2 dataset for natural language visual reasoning [35] as our source of images. We choose NLVR^2 for several reasons. First, it contains images of real-world entities with reasonable complexity in ImageNet-type [22]. Second, the setup of our task requires image in pairs that are similar enough, and NLVR^2 is designed to have collections of similar images regarding 1,000 synsets (*e.g.*, acorn, seawall). Also, Suhr et al. [35] employs an additional step to manually remove non-interesting images, thus ensuring the content quality.

Image subset construction. The nature of our task requires collections of negative images with high visual similarity, as otherwise, it would be trivial to discriminate between the reference and target image. Thus, prior to forming reference-target image pairs, we construct multiple subsets of six images that are semantically and visually similar, denoted as $\mathcal{S} = \{I_1, \ldots, I_6\}$, shown in Fig. 3(a).

Here, to construct a subset, we randomly pick one image from the large corpus $I_1 \in \mathcal{D}$. We then sort the remaining images in $\mathcal{D}$ by their cosine similarity to I_1 using ResNet152 [14] image feature vectors pre-trained on ImageNet [22]. Denote by κ_i the cosine similarity for image I_i. We then pick five additional images to produce a similar yet diverse subset, as follows: First, we filter out images with $\kappa_i \geq 0.94$ to avoid near-identical images to I_1. Then for the next top-20 ranked images, we greedily add each image in turn, skipping an image if its cosine similarity is within 0.002 of the last image added. If a subset of size six cannot be created, then the entire set is discarded.

Once constructed we further filter the collection subsets to avoid heavy overlap. We obtain in total 52,732 subsets from NLVR^2, from which we randomly choose 4,351 for the construction of CIRR.

Image pairing. Within each constructed image subset $\mathcal{S}$, we draw nine pairs of images, as shown in Fig. 3(b). We choose these pairs to have (1) consecutive modifications that will allow future training of a dialogue systems; and (2) multiple outcomes from the same reference image.

Annotations. We collect a modification sentence for each pair of reference-target images using Amazon Mechanical Turk (AMT). To ensure that no false-negatives exist within the same image subset from which we draw the pair, as illustrated in Fig. 3(c), we show AMT workers the remaining images from the subset and specifically ask them to write sentences that can *only* lead to the true target image.

AMT workers were instructed to avoid subjective descriptions, text mentions, plain side-by-side comparisons, or simple descriptions that only address the target images.

Following the collection of the modification sentences for each pair, we additionally collect some auxiliary annotations that more explicitly address the ambiguities associated with implicit human-agreements. While we believe that these auxiliary annotations will be useful for future work, we do not make use them in our current work[1].

Data splits. Following convention, we randomly assign 80% of the data for training, 10% for validation and 10% for test. Detailed statistics are shown in Table 2.

4.2. Analysis on CIRR

We follow Suhr *et al.* [35] and analyze coverage of various semantic concepts by keywords and sentence patterns (see Table 1). Here, we show comparisons with Fashion-IQ [12], the most popular, comparable human-labeled dataset. We observe a greater diversity and average length in the sentences in CIRR, indicating broad coverage and linguistic diversity. Over 40% of the annotations are compositional, which indicates an appreciable level of complexity of the sentences. Interestingly, our annotations should also encourage models to attend to both the reference and target images by implicitly (rows 1–4) or explicitly (rows 5–6) referring to the visual contents of *both* images.

5. Experiments

Datasets. To demonstrate the model's ability in untilizing pre-trained V&L knowledge, as well as its generalizability to images of different domains, we evaluate our proposed model against baselines and state-of-the-art (SoTA) methods on two datasets, including **(1)** CIRR, our proposed dataset on open-domain composed image retrieval, and **(2)** Fashion-IQ [12], which contains images of fashion products among three subtypes (`Dress`, `Shirt`, `Toptee`) with human-generated annotations. We do not evaluate on other datasets discussed in Sec. 2, as they either contain synthetic image/annotation or are domain-wise similar to Fashion-IQ (*e.g.*, Fashion200k [13]).

Compared methods. For CIRR, we evaluate the following methods using publicly available implementations[2]:

- TIRG [40] is an image-text composition model for composed image retrieval, which has proven to be effective on multiple datasets [12, 13, 18, 40]. The method uses a gating and residual design to encourage the learning of cross-modal features. Two setups for TIRG are available based on whether to inject text features at the last FC-layer (**default**), or the last convolution layer (**LastConv**). We test both setups.
- MAAF [8] is specifically designed for composed image retrieval with state-of-the-art performance. By default, it treats the convolutional spatial image features and the learned text embeddings (randomly initialized with LSTM [15]) as modality-agnostic tokens, which are passed to a Transformer [39]. We evaluate three design choices that were originally reported with comparable results: **(+BERT)** pretrained context-aware word representations using BERT [7], **(-IT)** removing the output of text tokens in the last pooling layer, **(-RP)** substituting the final resolution-wise pooling with average pooling.

[1] See supp. mat. and our project website for details on auxiliary annotations.

[2] `https://github.com/google/tirg`, `https://github.com/yahoo/maaf`

	Semantic aspect	Coverage (%)		Example (boldface added here for emphasis)
		CIRR	Fashion-IQ	
1	Cardinality	29.3	–	Only **one** of the boars and the ground is browner.
2	Addition	15.2	15.7	**Add** human feet and a collar.
3	Negation	11.9	4.0†	**Remove** the chair, make the dog sit in an open box.
4	Direct Addressing	57.4	49.0†	Show some lemons with a glass of lemonade.
5	Compare & Change	31.7	3.0	**Same computer but** different finish and black background.
6	Comparative Statement	51.7	32.0†	A **bigger** pub with **more** people on it.
7	Statement with Conjunction	43.7	19.0†	Remove all but one bird **and** have it facing right **and** putting food in its mouth.
8	Spatial Relations & Background	61.4	–	Change the sky to blue color.
9	Viewpoint	12.7	–	Focus widely on all available cookies package.
	Avg. Sentence length (words)	11.3	5.3	

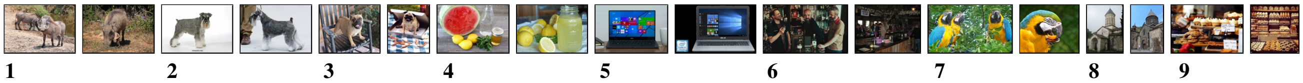


Table 1. Analysis of semantic aspects covered by the annotations in CIRR and in Fashion-IQ [12]. We also show average sentence length (nb. words). † Numbers from [12]. Image pair for each example is shown below with row number (left-right: reference-target).

	Nb. image subsets	Nb. pairs	Nb. pairs per subset	Nb. images
Train	3,345	28,225	7.54	16,939
Val.	503	4,184	8.32	2,297
Test	503	4,148	8.25	2,316
Total	4,351	36,554	8.40	21,552

Table 2. Statistics of CIRR. Each reference-target image pair is associated with one annotation.

For comparison, we also evaluate the following baselines, implemented by Vo et al. [40]:

- Random (theoretical): theoretical random guess.
- Random (init. ResNet): pretrained ImageNet [22] features, but random weights for others parameters.
- Image and text-only: substituting the combined image-text feature with the reference image or text feature.
- Random image with text: randomly sampling images to pair with text during training and validation.
- Concatenation: replacing the image-text composition layer with a simple concatenation of features followed by a 2-layer perceptron with ReLU.

For Fashion-IQ, we additionally include published results from the following methods:

- MRN [21] uses stacked blocks of element-wise products with residual learning to embed V&L jointly.
- FiLM [30] modulates the image feature map conditioned on text features after the layers of CNN.
- Relationship [33] learns the joint embeddings through relationship features constructed by concatenating the image and text features followed by FC-layers.
- VAL [5] is specially designed for composed image retrieval, which adopts the Transformer to compose multi-level V&L joint representations. For images with text descriptions as side information, an additional visual-semantic loss is applied to align visual features and the corresponding text features.

Metric. We follow previous work to report retrieval performance in Recall within top-K (Recall@K). For CIRR, we additionally report Recall$_{\text{subset}}$, which is an extension to the standard (global) Recall, made possible by the unique design of our dataset.

As discussed, our input queries $q = \langle I_{\text{R}}, t \rangle$ and target images I_{T} in our dataset are constructed such that both I_{R} and I_{T} are sampled from the same image set $\mathcal{S}$ (Sec. 4.1). We formulate Recall$_{\text{subset}}$ task by ranking images in $\mathcal{S} \setminus \{I_{\text{R}}\}$ according to model score. We define Recall$_{\text{subset}}$@K as the proportion of (test) examples where the ground-truth target image I_{T} is ranked within the top-K image in its subset.

Conceptually, Recall$_{\text{subset}}$ can be viewed as Recall while only considering images within the same subset as the pair. The benefits are twofold: First, Recall$_{\text{subset}}$ is not affected by false-negative samples, thanks to our careful design in data collection procedures. Second, with a selected batch of negative samples with high visual similarities, Recall$_{\text{subset}}$ can facilitate analysis on the reasoning ability of the methods for capturing fine-grained image-text modifications.

Implementation details. All experiments are conducted on a single NVIDIA RTX3090 with PyTorch. SoTA models use the default configurations proposed by their authors. See supp. mat. and our project website for more details on baseline training. For our proposed model, we use ResNet152 for image feature extraction. The model is optimized with AdamW [27] with an initial learning rate of 10^{-5}. We set a linearly decreasing schedule without warm-up. The batch size is set to 32 and the network is trained for 300 epochs. Other settings are kept as default by OSCAR.

5.1. Results

Baseline comparison on CIRR. Table 3 (rows 1-13) compares the retrieval performance of baseline and SoTA methods for both Recall and Recall$_{\text{Subset}}$@K on CIRR.

For global Recall, we notice that TIRG performs similar

		Methods	Recall@K				Recall$_{\text{Subset}}$@K			(R@5 + R$_{\text{Subset}}$@1)/2
			$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$	
BASELINES	1	Random (theoretical)	0.02	0.12	0.24	1.20	20.00	40.00	60.00	10.06
	2	Random (init. ResNet)	7.18	25.74	36.91	66.68	20.84	41.02	61.65	23.29
	3	Image-only	13.73	**48.46**	**65.81**	89.94	20.93	42.15	63.26	34.70
	4	Text-only	3.90	13.17	20.43	49.16	**39.69**	62.23	78.52	26.43
	5	Random Image+Text	2.99	11.91	19.85	46.97	39.41	**62.33**	**78.71**	25.66
	6	Image+Text Concatenation	12.44	40.24	57.52	87.29	23.74	45.12	65.50	31.99
	7	Human Performance†	–	–	–	–	86.09	–	–	–
SoTA	8	TIRG [40]	14.61	48.37	64.08	**90.03**	22.67	44.97	65.14	35.52
	9	TIRG+LastConv [40]	11.04	35.68	51.27	83.29	23.82	45.65	64.55	29.75
	10	MAAF [8]	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
	11	MAAF+BERT [8]	10.12	33.10	48.01	80.57	22.04	42.41	62.14	27.57
	12	MAAF−IT [8]	9.90	32.86	48.83	80.27	21.17	42.04	60.91	27.02
	13	MAAF−RP [8]	10.22	33.32	48.68	81.84	21.41	42.17	61.60	27.37
	14	Ours (no init.)	**15.18**	43.36	60.48	87.64	33.81	56.99	75.40	**38.59**
	15	Ours (init. OSCAR)	**19.55**	**52.55**	**68.39**	**92.38**	**39.20**	**63.03**	**79.49**	**45.88**

Table 3. Retrieval performance on CIRR. Best (resp. second-best) numbers are in bold-black (resp. blue). † See supplementary material on our collection details of human performance. We additionally report the average score over R@5 and R$_{\text{Subset}}$@1, which better reveals the overall performance of models (discussed in Sec. 5.1). Note that R@5 accounts for possible false-negatives in the entire image corpus. Since R$_{\text{Subset}}$ is not affected by such issues (Sec. 5), we consider R$_{\text{Subset}}$@1 to better illustrate the fine-grained reasoning ability of methods.

	Methods	Dress		Shirt		Toptee		Avg		(R@10 + R@50)/2
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
1	Image-only	4.20	13.29	4.51	14.47	4.13	14.30	4.28	14.20	9.15
2	Image+Text Concatenation	10.52	28.98	13.44	34.60	11.36	30.42	11.77	31.33	21.55
3	TIRG [40]	8.10	23.27	11.06	28.08	7.71	23.44	8.96	24.93	16.95
4	TIRG+Side Information [12]	11.24	32.39	13.73	37.03	13.52	34.73	12.82	34.72	23.77
5	MRN [21]	12.32	32.18	15.88	34.33	18.11	36.33	15.44	34.28	24.86
6	FiLM [30]	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04	25.28
7	TIRG [40]	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	27.40
8	Relationship [33]	15.44	38.08	18.33	38.63	21.10	44.77	18.29	40.49	29.39
9	VAL (init. GloVe) [5]	22.53	44.00	22.38	44.15	27.53	51.68	24.15	46.61	35.40
10	MAAF [8]	**23.8**	**48.6**	**21.3**	**44.2**	**27.9**	**53.6**	**24.3**	**48.8**	**36.6**
13	Ours (no init.)	14.38	34.66	13.64	33.56	16.44	38.34	14.82	35.52	25.17
14	Ours (init. OSCAR)	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20

Table 4. Retrieval performance on Fashion-IQ, we follow [12] to report average scores of R@10 and 50. Best numbers for SoTA models are in bold-black. Rows 1-4 reported by [12], rows 5-9 (shaded) reported by [5]. Rows 9-10 are SoTA methods developed for composed image retrieval, where we report the originally published numbers of their best configurations. Note that we see multiple scores reported for TIRG on Fashion-IQ, here we only show the published results from the above two sources. Additional non peer-reviewed methods that involve ensembles of models or data augmentation are not included.

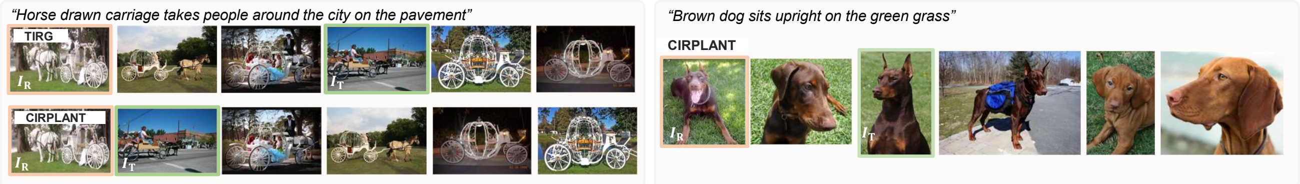

"Horse drawn carriage takes people around the city on the pavement"
TIRG
I_R
I_T
CIRPLANT
I_R
I_T
"Brown dog sits upright on the green grass"
CIRPLANT
I_R
I_T

Figure 4. Qualitative results of image retrieval on CIRR, red/green boxes: reference/target images. Predictions are ranked from left to right. We show the ranked images within subsets, see Sec. 5 for details on metric. (Left) We compare the retrieval on the same query for TIRG and CIRPLANT. (Right) We demonstrate the implicit ambiguities within the dataset (in this case, the difficulty in selecting the *most suitable* candidate by preserving the breed of the dog across the images, which requires identifying subtle characteristics– *e.g.* pointy ears).

to the Image-only baseline, suggesting that its multi-modal composition layers often fail to extract information from the text. Instead, it relies primarily on visual content. We conjecture that CIRR focuses more on the fine-grained changes that are harder to capture and associate across modalities, therefore, requires stronger image-text composition layers. In addition, we note that MAAF (rows 10-13) does not generalize well to our dataset, even though it outperforms TIRG and other methods on existing ones [8]. We believe the choice of forming image tokens by spatial feature maps does not generalize to our dataset where the modification concepts are more diverse and at multiple levels. Meanwhile, adding the contextual-aware BERT pretrained weights yields little effects, suggesting a plain initialization of word embeddings, though contains validated pre-trained language information, may not help the composition layers.

The Recall$_{\text{Subset}}$ results tell a similar story. Here the performance of all SoTA models is close to the theoretical random guess, indicating that current models fail to capture fine-grained modifications between similar images. Interestingly, we discover that the Text-only and Random-Image+Text baselines (rows 4,5) outperform SoTA models significantly. We believe this is because the modification sentences usually contain descriptions of visual content that is unique to the target image once limited to the smaller retrieval set (*e.g.*, "add a leash to the dog" where only the target image contains the leash). However, as demonstrated by the low Recall performance, such descriptions are not detailed enough to single out the target image in the entire image corpus. This scenario further demonstrates Recall$_{\text{Subset}}$ reveals behaviors of models on different aspects, and can be used for more detailed analysis.

In short, the relatively low retrieval performance suggests that our dataset poses a challenge to existing methods developed and tested on narrow-domain datasets.

Performance of CIRPLANT on CIRR. Results in Table 3 (rows 14,15) compares our proposed model with SoTA methods on CIRR. We notice that on CIRR, CIRPLANT with no initialization (row 14) performs similarly as TIRG on Recall, while surpassing all other SoTA methods. This validates our design choice of using non-regional image features for composing image and text through the transformer architecture. Meanwhile, on Recall$_{\text{Subset}}$ our model, even without initialization, yields much higher scores than others, suggesting transformers are better in capturing more fine-grained visiolinguistic cues when composing image and text features. Comparing with SoTA methods that use LSTMs for generating a single language embedding of the entire sentence, we believe that the key difference lies within the fact that transformers accept word tokens as input, which can later be attended individually. Our model outperforms all other methods with OSCAR initialization (row 15) by a significant margin, demonstrating the benefit of VLP knowledge on open-domain images.

Performance of CIRPLANT on Fashion-IQ. Table 4 compares the performance of our model with SoTA methods. We notice that our model with OSCAR initialization (row 14) outperforms most methods, including generic multimodal learning methods and TIRG. This strengthens the benefits of using transformer architecture that leverages VLP models. Additionally, we note that even on Fashion-IQ, our model still benefits greatly from OSCAR pre-trained initialization (rows 13,14). Given that the images in Fashion-IQ differ greatly from the data used for OSCAR pre-training [25], we believe this further demonstrates that the pre-trained model can transfer the learned V&L knowledge and adapt to various contexts.

We note that two recent SoTA methods for composed image retrieval (VAL and MAAF, rows 9,10) perform better than our model. Despite the visible improvements brought by OSCAR initialization, we hypothesize that our model is still underperformed by the apparent domain shift in images, as the VLP model is pre-trained on generic ImageNet-type data. Meanwhile, the low generalizability of MAAF on CIRR (Table 3 rows 10-13) hints the possibility that current SoTA methods developed and tested on existing datasets may have been overly adapted to domain-specific images of low complexity. Hence, additional open-domain datasets, such as CIRR, can be beneficial in future research.

5.2. Qualitative Results

Fig. 4 (left) demonstrates the retrieval rankings within the image subset (see Sec. 5) on the same query for TIRG and CIRPLANT. Specifically, we show the effectiveness of pre-training in CIRPLANT when encountering visiolinguistic concepts (*i.e.*, *pavement*) that occur less frequently in the training data. Additionally, CIRPLANT better captures fine-grained cues within language (*e.g.*, *takes people around*, which implies *must have people in the back of the carriage*), thanks to the transformer architecture that accepts, and attends to individual word tokens.

We show one failure case of CIRPLANT on CIRR in Fig. 4 (right). Note the implicit requirement of *preserving same breed of dog* across the reference and target image. This requires models to identify the fine-grained visiolinguistic cues (*i.e.*, pointy ears in this sample) and retrieve the most suitable image, bringing more challenge to the task.

6. Conclusion

This work expands the task of composed image retrieval into more complex, open-domain images. We collect the CIRR dataset, which addresses shortcomings of existing datasets by placing more emphasis on distinguishing open-domain visually similar images. Our publicly available dataset is designed to facilitate future studies on subtle reasoning over visiolinguistic concepts, as well as iterative retrieval with dialogue. We also introduce CIRPLANT, a transformer-based model that leverages V&L pre-training to compose image and text features. We validate CIRPLANT on both CIRR and the existing fashion dataset, demonstrating the generalizability of our design and the effectiveness of V&L pre-training. Collectively, we hope to inspire future work on composed image retrieval on a broader scope, yet fine-grained level.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. 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

[2] J. Aneja, A. Deshpande, and A. G. Schwing. Convolutional image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision*, 2015. 2

[4] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, 2010. 1, 3

[5] Y. Chen, S. Gong, and L. Bazzani. Image search with text feedback by visiolinguistic attention learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 6, 7

[6] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 2020. 2

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019. 2, 3, 5

[8] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, and K. Boakye. Modality-agnostic attention fusion for visual search with text feedback. *ArXiv*, abs/2007.00145, 2020. 2, 3, 5, 7

[9] M. Forbes, C. Kaeser-Chen, P. Sharma, and S. J. Belongie. Neural Naturalist: Generating fine-grained image comparisons. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 3

[10] R. B. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, 2015. 2

[11] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, 2018. 3

[12] X. Guo, H. Wu, Y. Gao, S. J. Rennie, and R. Feris. The Fashion IQ Dataset: Retrieving images by combining side information and relative natural language feedback. *ArXiv*, abs/1905.12794, 2019. 1, 2, 3, 4, 5, 6, 7

[13] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *IEEE International Conference on Computer Vision*, 2017. 1, 2, 3, 5

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 5

[15] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9:1735–1780, 1997. 5

[16] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould. A recurrent vision-and-language bert for navigation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[17] M. Hosseinzadeh and Y. Wang. Composed query image retrieval using locally bounded features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[18] P. Isola, J. J. Lim, and E. H. Adelson. Discovering states and transformations in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2, 3, 5

[19] H. Jhamtani and T. Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Conference on Empirical Methods in Natural Language Processing*, 2018. 3

[20] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[21] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Advances in neural information processing systems*, 2016. 6, 7

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Association for Computing Machinery*, 2017. 4, 5, 6

[23] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language, 2019. 2, 3

[24] W. Li, L. Duan, D. Xu, and I. W. Tsang. Text-based image retrieval using progressive multi-instance learning. In *IEEE International Conference on Computer Vision*, 2011. 1, 2

[25] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 2020. 2, 3, 8

[26] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[27] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[28] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019. 2

[29] I. Masi, Y. Wu, T. Hassner, and P. Natarajan. Deep face recognition: A survey. In *SIBGRAPI Conference on Graphics, Patterns and Images*, 2018. 2

[30] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer, 2017. 6, 7

[31] F. Radenović, G. Tolias, and O. Chum. Deep shape matching. In *European Conference on Computer Vision*, 2018. 2

[32] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 2

[33] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017. 6, 7

[34] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[35] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi. A corpus for reasoning about natural language grounded in photographs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 2, 4, 5

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

[37] D. Teney, P. Anderson, X. He, and A. V. D. Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[38] S. Tong and E. Chang. Support Vector Machine active learning for image retrieval. In *Proceedings of the Ninth ACM International Conference on Multimedia*, 2001. 1

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin. Attention is all you need. In *International Conference on Neural Information Processing Systems*, 2017. 2, 3, 5

[40] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3, 4, 5, 6, 7

[41] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI Conference on Artificial Intelligence*, 2019. 2

[42] C. Zhang, J. Y. Chai, and R. Jin. User term feedback in interactive text-based image retrieval. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005. 1, 2