

# Towards Fast and Accurate Image-Text Retrieval with Self-Supervised Fine-Grained Alignment

Jiamin Zhuang, Jing Yu\*, Yang Ding, Xiangyan Qu, Yue Hu

**Abstract**—Image-text retrieval requires the system to bridge the heterogenous gap between vision and language for accurate retrieval while keeping the network lightweight-enough for efficient retrieval. Existing trade-off solutions mainly study from the view of incorporating cross-modal interactions with the independent-embedding framework or leveraging stronger pre-trained encoders, which still demand time-consuming similarity measurement or heavyweight model structure in the retrieval stage. In this work, we propose an image-text alignment module SelfAlign on top of the independent-embedding framework, which improves the retrieval accuracy while maintains the retrieval efficiency without extra supervision. SelfAlign contains two collaborative sub-modules that force image-text alignment at both concept level and context level by self-supervised contrastive learning. It doesn't require cross-modal embedding interactions during training while maintaining independent image and text encoders during retrieval. With comparable time cost, SelfAlign consistently boosts the accuracy of state-of-the-art non-pre-training independent-embedding models respectively by 9.1%, 4.2% and 6.6% in terms of R@sum score on Flickr30K, MSCOCO 1K and MS-COCO 5K datasets. The retrieval accuracy also outperforms most existing interactive-embedding models with orders of magnitude decrease in retrieval time. The source code is available at: <https://github.com/Zjamie813/SelfAlign>.

**Index Terms**—Fast image-text retrieval, concept-level cross-modal alignment, context-level cross-modal alignment, self-supervised learning.

## I. INTRODUCTION

IMAGE-TEXT retrieval (ITR) is a long-standing task that requires an AI agent to retrieve semantically relevant images given a text query and vice versa. The key challenge of ITR is to bridge the heterogeneous gap between low-level visual appearance and high-level abstract language and align their representations. It is also a fundamental problem for a series of vision and language tasks [1, 20, 27]. In real-world scenarios, besides effective cross-modal alignment for accurate retrieval, the retrieval system also strives to make real-time retrieval possible with low latency. Therefore, how to balance the accuracy and efficiency becomes a key challenge for large-scale image-text retrieval.

Most of the previous works make much effort on either retrieval efficiency or retrieval accuracy. Early *independent-embedding approaches* [10, 23, 30] (Figure 1(a)) encode each

This work was supported by the National Natural Science Foundation of China (Grant No. 62006222) and the Youth Innovation Promotion Association of CAS (Grant No. 2021153).

Jiamin Zhuang, Jing Yu, Yang Ding, Xiangyan Qu and Yue Hu are with the Institute of Information Engineering, Chinese Academy of Sciences, China, and the School of Cyber Security, University of Chinese Academy of Sciences, China. (e-mail: zhuangjiamin@iie.ac.cn; yujing02@iie.ac.cn; dingyang@iie.ac.cn; quxiangyan@iie.ac.cn; huyue@iie.ac.cn)

Corresponding author: Jing Yu (e-mail: yujing02@iie.ac.cn)

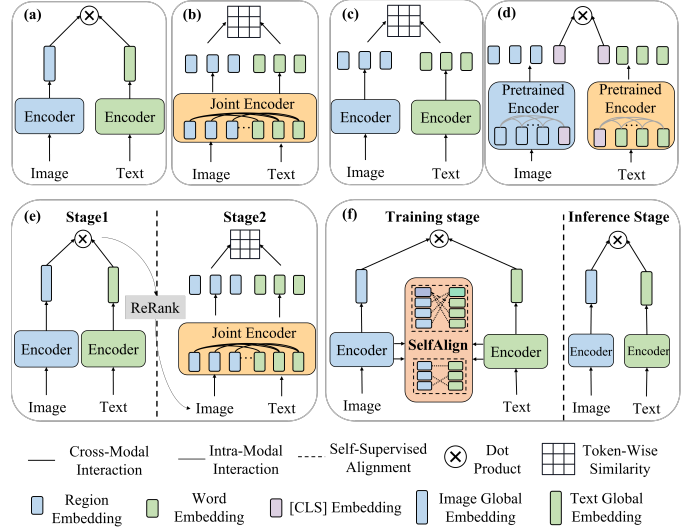


Fig. 1. Illustration of different image-text retrieval approaches. (a) Independent-embedding approach. (b) Interactive-embedding approach. (c) Late-interaction approach. (d) Intra-interactive embedding approach. (e) Two-stage approach. (f) Independent-embedding approach with SelfAlign.

image and text independently into global embeddings. Then image-text similarity is computed by directly measuring the distance between their global embeddings in a common semantic space. Since there are no interactions between texts and images, independent-embedding approaches allow offline data embedding extraction and linear computational complexity [26] for online retrieval. Hence such approaches are widely applied in real-world large-scale retrieval. However, their retrieval accuracy is not satisfactory since such a global embedding alignment strategy cannot guarantee fine-grained content alignment. To alleviate this problem, *interactive-embedding approaches* [5, 21, 31] (Figure 1(b)) are proposed for fine-grained image-text retrieval by aligning visual objects in an image with words in a text by cross-modal attention mechanism. However, for each query, all the retrieved samples need complex attention computation to encode their embeddings, which is quite time-consuming and not scalable to large-scale online retrieval scenarios. *How to leverage the advantages of independent-embedding approaches and interactive-embedding approaches to achieve both high accuracy and practical efficiency becomes an essential problem.*

Current progress [12, 28, 32] aims to introduce computation-efficient interactions into the independent-embedding framework. The typical solutions can be divided into three types: measuring fine-grained word-object similarities instead of global embedding similarities

(a.k.a. *late-interaction approach* as shown in Figure 1 (c)) [26, 28], adopting independent-embedding approaches for coarse retrieval first and then using interactive-embedding approaches for finer retrieval (a.k.a. *two-stage approach* as shown in Figure 1 (e)) [12, 29], and exploiting stronger intra-modal interactive encoder instead of the time-consuming cross-modal interactive encoder (a.k.a. *intra-interactive embedding approach* as shown in Figure 1 (d)) [17, 32]. However, compared with independent-embedding approaches, these trade-off solutions still demand extra time cost in the retrieval stage due to complex similarity measurement, cross-modal embedding interactions, or heavyweight encoder structure. They sacrifice the retrieval efficiency for the benefits of fine-grained feature learning.

In this paper, to enable fine-grained image-text alignment for accurate retrieval while maintain high retrieval efficiency as the independent-embedding models, we propose a novel trade-off strategy to learn fine-grained image-text alignment by contrastive-based embedding mapping. The advantage of contrastive-based embedding mapping is that it does not require cross-modal fusion during the inference stage. In our approach, we achieve embedding alignment between images and texts by a new module, named as SelfAlign. Based on the backbone of independent-embedding models, SelfAlign aims to align image and text embeddings from local to global by multi-level self-supervised contrastive learning. In this way, independent-embedding models injected with SelfAlign enhance original global embedding with more accurate fine-grained semantic alignment across different modalities. In the inference stage, the baseline model without the SelfAlign module conducts image and text encoding independently while maintains fine-grained embedding alignment ability. Therefore, SelfAlign enables independent-embedding models to achieve superior retrieval accuracy without sacrificing efficiency.

Specifically, SelfAlign is designed to explore fine-grained correspondence via mining rich visual and textual semantic content in different layers of the independent-embedding models. There are two sub-modules in SelfAlign responsible for semantic alignment at concept level and context level: (1) To capture the pair-wise correspondence among visual region and textual words, Local Concept Alignment (LCA) sub-module is first proposed to learn the concept-level alignment in the lower layer. (2) Since the semantically similar concepts have different semantics in different contexts, Context Relation Alignment (CRA) sub-module is further proposed to be injected into higher embedding layer to achieve context-level alignment. As a result, independent-embedding models with SelfAlign learn fine-grained alignment between images and texts from local semantics to global semantics progressively.

The main contributions are summarized as follows: (1) We introduce a novel trade-off strategy for image-text retrieval to learn fine-grained alignment by contrastive-based embedding mapping. The contrastive-based embedding mapping aligns the fine-grained image and text embeddings via cross-modal contrastive learning during the training stage without requiring cross-modal fusion during the inference stage. Thus, our approach has the benefits of both the interactive-embedding

models and the independent-embedding models, i.e., enhancing fine-grained alignment learning while preserving the independent-embedding framework for efficient retrieval. (2) We propose a fine-grained image-text alignment module SelfAlign to achieve the trade-off strategy. Two sub-modules of SelfAlign conduct concept alignment and context alignment via cluster-based contrastive learning and global-to-local contrastive learning. Therefore, SelfAlign equips the global embedding of the independent-embedding models with multi-level fine-grained alignment to improve the retrieval accuracy without extra supervision. (3) SelfAlign is a generic module that can be injected into various independent-embedding models. We incorporate SelfAlign with two representative independent-embedding models. Experimental results show that SelfAlign consistently boosts the accuracy of state-of-the-art independent-embedding models respectively by 9.1%, 4.2% and 6.6% in terms of R@sum on Flickr30K, MS-COCO 1K and MS-COCO 5K. The performance also outperforms most existing interactive-embedding models with orders of magnitude decrease of retrieval time.

## II. RELATED WORKS

### A. Image-Text Retrieval

Existing works can be categorized into two types: the independent-embedding approaches and the interactive-embedding approaches. The former approaches [10, 11, 23, 30, 45] aim to project the images and texts into a common semantic space, so image-text pairs can be compared directly via simple distance metrics. The architecture of mainstream independent-embedding approaches is an independent-embedding learning structure consisting of an image encoder and a text encoder, and they adopt ranking loss [10] for metric learning. Though these approaches have achieved some promising performance, they are still limited since they cannot conduct interactive encoding process and thus fail to provide fine-grained alignment between images and texts.

Interactive-embedding approaches [5, 16, 18, 21, 25, 31] aim to learn fine-grained image-text matching by complex object and word interactions with cross-modal attention mechanism. Lee et al. [21] compute the similarities between regions and words, and only count the region-word pairs with high relevance. Some works [5, 16, 25] propose hierarchical interaction methods for progressively extracting the complicated correspondence. Recently, Qu et al. [31] propose a dynamic router with the capability to choose the different interactive mode for each image-text pair and achieves state-of-the-art performance. Nevertheless, the quadratic computational complexity takes unavoidable computational cost for retrieval. In this paper, we propose a model-agnostic module with multi-level self-supervised learning strategy for independent-embedding models to learn the fine-grained semantic correspondences, instead of time-consuming attention mechanisms. Thus, the independent-embedding backbones achieve superior retrieval accuracy without sacrificing efficiency.

### B. Trade-Off Image-Text Retrieval Models

To strike a balance between retrieval efficiency and accuracy, there are almost three types of approaches that have

been proposed recently: late-interaction approaches [26, 28], two-stage approaches [12, 29], and intra-interactive embedding approaches [17, 34]. The late-interaction approaches retain the independent encoding architecture and perform lightweight token-wise interactions only in the late scoring stage [26, 28]. Their retrieval speed is still slower than independent-embedding methods since the independent-embedding methods only require the global embedding for similarity computation. Secondly, two-stage approaches [12, 29] first adopt independent-embedding models for coarse-level retrieval and then utilize interactive-embedding models for finer retrieval to trade-off between efficiency and accuracy. But they are still slower than independent-embedding models due to the existence of the re-rank stage. Lastly, the intra-interactive embedding approaches [17, 34] take two independent encoders but they require large-scale image-text pairs for training and stack a few Transformer blocks [37] to build stronger encoders. For example, ALIGN [17] leverages two transformers encoders with 400M parameters and 1.8B image-text pairs for training. Though they achieve surprising performance, the inevitable huge computation cost in training and deploying such massive-scale models limit their development. Contrast to these three types of trade-off methods, without sacrificing the retrieval time or requiring extra training data, our proposed module improves the accuracy by enhancing image and text representations of fine-grained information.

### C. Self-Supervised Contrastive Learning

Self-supervised learning [6, 35, 46] aims at learning features without manual annotations. Recent approaches based on contrastive learning have achieved remarkable progress in visual domain. Current contrastive learning can be divided into two groups: individual-based contrastive learning [6, 13] and cluster-based contrastive learning [3, 4]. Individual-based contrastive learning [6, 13] considers each image in a dataset as its own class [4], and brings the embedding of different views from the same image closer and push embeddings from different images far apart using instance-level contrastive loss. This approach introduces the individual-level discrimination but requires a large batch size for negatives storage. Cluster-based contrastive learning [3, 4] encourages the image embeddings to be closer to their assigned prototypes obtained by clustering algorithm, and far from negative prototypes. This approach introduces the group-level discrimination between instances. However, current works only construct image-level representation learning and lack of fine-grained information learning such as the object information in images.

To learn local information, some researchers [22, 41] construct local level contrastive learning to learn visual pixel-level semantic information. Moreover, some works [2, 14] maximize global-local mutual information and aim to learn the shared context information across patches/tokens. Here, we aim to learn the fine-grained correspondences between images and texts without detail annotations, and our word-object alignment learning and context-level alignment learning were inspired by these local contrastive learning works. Differently, the contrastive reasoning should be performed across modalities instead of cross-image views, and thus the learning process

is not fully symmetric since the local semantic information involved in image-text pairs is not exactly equivalent to that of two augmentations of an image.

## III. METHODOLOGY

We propose a module SelfAlign to explore the multi-grained correspondences in the different layers of independent-embedding models. Independent-embedding models mainly consist of a visual encoder and a textual encoder, as shown in Figure 1 (a). Though the encoders are various from adopting different encoding architectures such as GCN [19] and self-attention [30, 45], the encoding process for each modality typically includes three stages as shown in the Figure 2 (a), visual object encoding and textual word encoding stage, visual and textual context encoding stage and visual and textual embedding aggregation stage. However, independent-embedding models only conduct global image-text alignment at the embedding aggregation stage and overlook fine-grained alignment in the first two encoding stages. To improve their retrieval accuracy, we design two sub-modules in SelfAlign injected in the first two encoding stages respectively: 1) **Local Concept Alignment (LCA)** sub-module for local conceptual level alignment between visual objects and textual words, 2) **Contextual Relation Alignment (CRA)** sub-module for contextual level alignment.

In this section, we first describe the single modal embedding extraction approaches in the independent-embedding models in Section III-A. In Section III-B, we introduce the LCA sub-module, which learns the concept-level word-object correspondences at the object and word encoding stage. We then introduce the CRA sub-module in Section III-C to explore context-level alignment in the context encoding stage. Since SelfAlign is model-agnostic and applicable to independent-embedding models, we case study on two representative baseline models, VSRN [23] and CAMERA [31], which is introduced in Section III-D and Section III-E, respectively.

### A. Single-Modal Embedding Extraction

**Image Embedding Extraction.** For each input image  $I$ , recent works [21, 23] usually employ an off-the-shelf object detection model, such as Faster R-CNN [33], to detect  $M$  objects  $O = \{o_i\}_{i=1}^M$ , where each object  $o_i$  is represented by an object feature embedding  $\mathbf{o}_i \in \mathbb{R}^{d_o}$ . Then a linear projection is utilized to transform  $\mathbf{o}_i$  into a  $h$ -dimensional embedding. Then the embedding of the image is represented by a set of object embeddings  $\mathbf{V}^l = \{\mathbf{v}_i^l\}_{i=1}^M$ . We name this encoding process as visual object encoding stage. Then different works utilize various approaches such as GCN [23] or Transformers [23] to model the relationships between objects and obtain contextualized object embeddings, named as visual context encoding stage. Here, we omit the computation details and simplify them as the visual context encoder, and the output of the context encoder is denoted as visual context embeddings  $\mathbf{V}^c = \{\mathbf{v}_i^c\}_{i=1}^M$ . Finally, the visual global embedding  $\mathbf{V}^g$  is obtained by integrating the context embeddings  $\mathbf{V}^c$ , which is named as visual embedding aggregation stage.

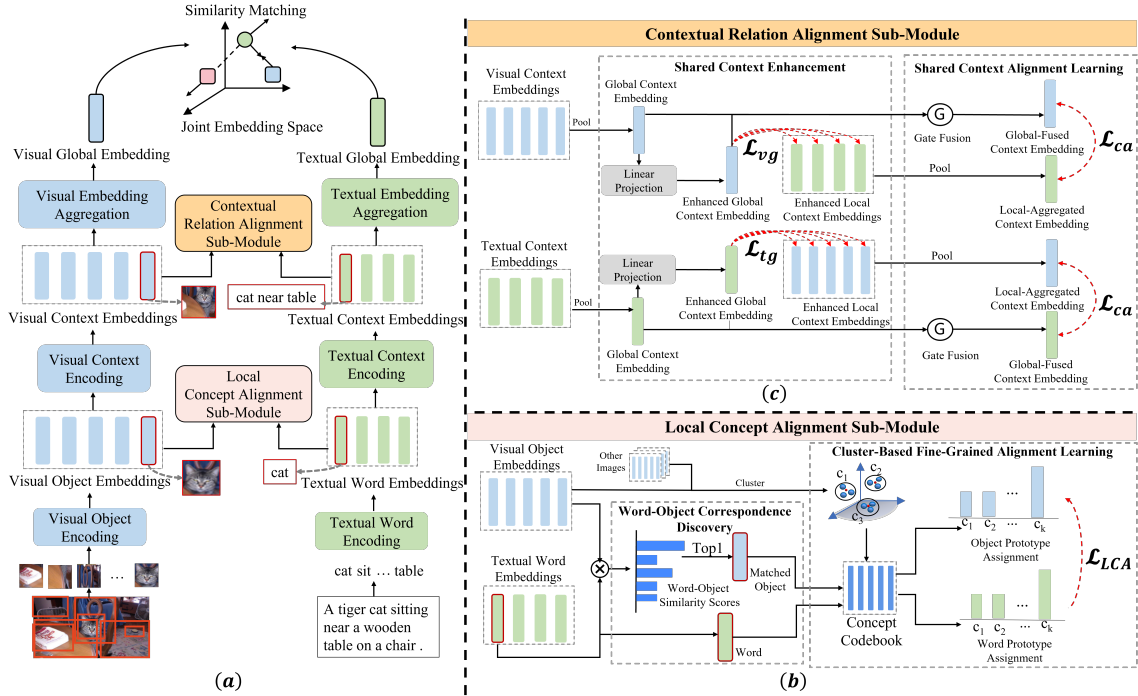


Fig. 2. The overview of SelfAlign applied to independent-embedding models. SelfAlign consists of two sub-modules: Local Concept Alignment sub-module and Contextual Relation Alignment sub-module. (a) illustrates the hierarchical encoding stage of independent-embedding models and the inject position for the two sub-modules of SelfAlign. (b) describes the Local Concept Alignment sub-module and (c) describes the Contextual Relation Alignment sub-module.

**Text Embedding Extraction.** For each input text  $T$ , the word-level embeddings  $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^D$  is generally obtained from the pre-trained word embeddings [21, 23], such as BERT [8]. Then an embedding layer is also employed to transform  $\mathbf{w}_i$  to a  $h$ -dimensional vector. The word embedding set is denoted as  $\mathbf{T}^l = \{\mathbf{t}_i^l\}_{i=1}^D$ . We name this encoding process as the textual word encoding stage. Then different works utilize different approaches, such as GRU [21, 23] and Transformer [17, 30], to capture the contextualized word embeddings, named as the textual context encoding stage. We denote the output of this stage as textual context embeddings  $\mathbf{T}^c = \{\mathbf{t}_i^c\}_{i=1}^D$ , where  $D$  denotes the number of the words in the sentence. The final global embedding  $\mathbf{T}^g$  is obtained by aggregating the context embeddings  $\mathbf{T}^c$ , named as textual embedding aggregation stage.

### B. Local Concept Alignment Sub-Module

Local Concept Alignment (LCA) sub-module is injected into the object and word encoding layer of the baseline model and aims to force the consistency between the visual and textual concept embeddings, as presented in Figure 2 (b). We regard the object regions as the visual concepts and the whole words in the sentences as textual concepts to make full use of information in the whole sentence. Since there are no word-object pair annotations, a word-object correspondence discovery strategy is first proposed to build pseudo word-object correspondences in LCA. Then a cluster-based fine-grained alignment is designed to force the consistency between the selected matched word-object pair by comparing their cluster assignments. Though there are no annotations of the word-

object pairs, our concept alignment is under the constraint of global image-text matching supervision.

**Word-Object Correspondence Discovery.** We take the visual object embeddings  $\mathbf{V}^l$  and word embeddings  $\mathbf{T}^l$  as the input of the LCA sub-module. Since there are no word-object annotations, we first estimate the word-object correspondences from image-text pairs. Specifically, we compute the cosine similarities between the words and objects and choose the most similar object as the correspondence for each word. Notably, since the texts of an image is the linguistic expression of the image itself [24, 48], each word is able to find at least one matched object region given an image-text pair. We cannot guarantee each object in the image has corresponding words in the text. Therefore, we align each word to an object rather than the inverse direction.

Formally, given the  $i$ -th word embedding  $\mathbf{t}_i^l$ , we compute its cosine similarities with the region embeddings  $\mathbf{V}^l$  and apply the argmax operation to select the most matched object region:

$$\mathbf{v}_{j^+}^l = \arg \max_{j \in [1, M]} \cos(\mathbf{t}_i^l, \mathbf{v}_j^l) \quad (1)$$

where  $\mathbf{v}_{j^+}$  is the selected matched image object region for  $\mathbf{t}_i$  and  $\cos(\cdot, \cdot)$  represents the cosine similarity function. We utilize the whole word-object pairs obtained by all words in a batch as pseudo-supervised alignment information for local concept-level alignment.

**Cluster-based Fine-grained Alignment.** Cluster-based fine-grained alignment aims to enforce the consistency between each word embedding and the matched object embedding selected by the word-object correspondence discovery strategy. Inspired by the recent success of self-supervised



learning [6, 35], we treat the selected matched word-object pair as two different views depicting the same conceptual semantic and utilize cluster-based contrastive learning [3, 4] to align their embeddings across different modalities. The basic idea lies in that we cluster object region embeddings to form a concept dictionary, where each cluster center represents a concept. The embeddings of the cluster centers are more stable and representative to represent concepts compared with the object region embeddings in each image. Each region embedding and word embedding are then mapped to the concept dictionary and represented by concept assignments. In this way, we represent regions and words at the semantic-consistent concept level instead of the expression-variant instance level. After that, we utilize cross-prediction to align the region and word embeddings. Notably, there is an alternative solution to align the embeddings, named as individual-based contrastive learning [6, 13]. However, individual-based contrastive learning requires a large number of paired negative-positive samples, and it is challenging to select the appropriate negative samples without ground-truth word-object annotations. Thus, to efficiently and effectively align the word-object embeddings, we adopt cluster-based contrastive learning without requiring numerous computations. It is an alternative to contrasting multiple views by comparing their cluster assignments instead of their instance-level embeddings [4].

The cluster-based fine-grained alignment process contains two main steps: (1) We first perform *concept codebook construction* to obtain the concept assignments, where the concept codebook is a collection of learnable prototypes based on object region embeddings. (2) Then *cross-prediction learning* is conducted from word concept assignments to object concept assignments to align the word-object embeddings.

*Step 1: concept codebook construction.* To learn the concept codebook and let the learning process of cluster centers be synchronous with the whole network, we use the online clustering method as in [4]. Following [4], we define the trainable concept codebook  $\mathcal{C} = \{c_i\}_{i=1}^K$ , where  $K$  represents the number of concept centers and each center  $c_i$  is a  $h$ -dimensional vector as visual and textual concept embeddings.

*Step 2: cross-prediction learning.* We use cross-entropy loss between object and word concept assignments to set up this cross-prediction problem. Firstly, given the concept codebook, the visual concept assignment  $v_i^P \in \mathbb{R}^K$  for the object embedding  $v_i^l$  can be obtained by mapping  $v_i^l$  to the concept codebook. Each scalar  $p_i^k$  in  $v_i^P$  represents the probability that  $v_i$  belongs to the  $k$ -th prototype  $c_k$  and is obtained by taking a softmax of the cosine similarity of  $v_i^l$  and  $c_k$  as follows:

$$p_i^k = \frac{\exp(\cos(v_i^l, c_k) / \tau_1)}{\sum_{k=1}^K \exp(\cos(v_i^l, c_k) / \tau_1)} \quad (2)$$

where  $\tau_1$  is a temperature parameter. Similarly, the word assignment is computed by mapping each word embedding  $t_i^l$  to the concept codebook, and each scalar of the probability that  $t_i^l$  belongs to the  $k$ -th prototype, denoted as  $q_i^k$ .

Then cross-prediction from word concept assignments to the corresponding object concept assignments is performed to align the selected word-object pair. The prediction problem is

optimized by the cross-entropy loss as follows:

$$\mathcal{L}_i = - \sum_{k=1}^K p_{j^+}^k \log q_i^k \quad (3)$$

where  $p_{j^+}^k$  is the object concept assignment of the matched visual object embedding  $v_{i^+}^l$ . Finally, we aggregate the loss values in Equation 3 over all the words  $D$  and result in the following loss function to optimize the feature encoders:

$$\mathcal{L}_{LCA} = \frac{1}{D} \sum_{i=1}^D \mathcal{L}_i \quad (4)$$

### C. Contextual Relation Alignment Sub-Module

Since semantically similar concepts have different semantics in different contexts, Context Relation Alignment (CRA) submodule is further proposed to capture context-level semantic correspondences in the context embedding layer, as presented in Figure 2 (c). CRA first performs shared context enhancement to capture the shared context-level information and suppress the irrelevant information between the images and texts. Shared context alignment is then conducted to achieve contextual-level alignment.

**Shared Context Enhancement.** Given the visual context embeddings  $V^c = \{v_i^c\}_{i=1}^M$  and text context embeddings  $T^c = \{t_i^c\}_{i=1}^D$ , we acquire the global context embedding  $v_g^c$  and  $t_g^c$  for contextual relation alignment by adopting the average-pooling over  $V^c$  and  $T^c$  respectively. Here, we design two symmetric contrastive mechanism to learn the shared context information from visual perspective and textual perspective: contrasting between the visual global context embedding and the textual local context embeddings, denoted as V-global/T-local and contrasting between the textual global context embedding and the visual local context embeddings, denoted as T-global/V-local. Under the global supervision from one modality, the relevant local semantics of objects and relationships in the other modality will be strengthened while irrelevant semantics will be weakened.

Formally, a fully connected layer is first utilized to map the global context embedding into the corresponding local context space as the global supervision:

$$\begin{aligned} t_s^c &= \sigma(\text{BN}(\mathbf{W}_1 t_g^c + b_1)) \\ v_s^c &= \sigma(\text{BN}(\mathbf{W}_2 v_g^c + b_2)) \end{aligned} \quad (5)$$

where  $\sigma$  is the ReLU activation and BN is the batch normalization.  $\mathbf{W}_1$ ,  $\mathbf{W}_2$ ,  $b_1$  and  $b_2$  are learnable parameters. Meanwhile, the batch normalization and ReLU activation operations are employed on the local contextual embeddings  $v_i^c$ ,  $t_i^c$ , and they are then denoted as  $v_i^*$ ,  $t_i^*$  in contrastive learning.

Given the global context and local context embeddings, the T-global/V-local contrastive learning and V-global/T-local contrastive learning is conducted in parallel. For T-global/V-local contrastive learning, we consider the textual global supervision embedding  $t_s^c$  and the visual local contextual embeddings  $v_i^*$  that come from an image-text pair as a positive sample, while the embeddings from unpaired image-text pairs as negative

samples. Then we define the T-global/V-local contrastive loss as:

$$\mathcal{L}_{tg} = -\frac{1}{M} \sum_{i=1}^M \log \frac{\exp(\cos(\mathbf{t}_s^c, \mathbf{v}_i^*)/\tau_2)}{\sum_{j=1}^N \exp(\cos(\mathbf{t}_s^c, \tilde{\mathbf{v}}_j^*)/\tau_2)} \quad (6)$$

where  $\tau_2$  is a temperature hyper-parameter and  $M$  is the number of visual regions.  $\{\tilde{\mathbf{v}}_j^*\}_{j=1}^N$  is a set of negative visual local embeddings.  $N$  is the number of negatives. Similarly, the V-global/T-local contrastive loss is defined as:

$$\mathcal{L}_{vg} = -\frac{1}{D} \sum_{i=1}^D \log \frac{\exp(\cos(\mathbf{v}_s^c, \mathbf{t}_i^*)/\tau_2)}{\sum_{j=1}^N \exp(\cos(\mathbf{v}_s^c, \tilde{\mathbf{t}}_j^*)/\tau_2)} \quad (7)$$

where  $\{\tilde{\mathbf{t}}_j^*\}_{j=1}^N$  is negative textually local context embeddings.  $D$  is the number of textual words. We select the  $N$  negative local embeddings that are most similar to the global embedding of the other modality ranked by cosine similarity. As a result, these samples are hard negative samples which are beneficial for learning high-quality representations of the anchor sample [47]. The loss for shared context enhancement is defined as:

$$\mathcal{L}_{cs} = \frac{1}{2} (\mathcal{L}_{tg} + \mathcal{L}_{vg}) \quad (8)$$

**Shared Context Alignment.** Shared context alignment aims to perform contextual level alignment based on the enhanced global context embedding and local context embeddings. Specifically, both of the V-global/T-local and the T-global/V-local contrastive learning enhance the global context embeddings  $\mathbf{v}_g^c$  and  $\mathbf{t}_g^c$  from local and global perspectives and introduce a paired final global context embeddings for context alignment. Taking T-global/V-local contrastive learning as an example, it enables the encoder to filter out irrelevant local contextual information which is not aligned with global contextual information of the other modality. We further use average-pooling on visual enhanced local context embeddings  $\{\mathbf{v}_i^*\}_{i=1}^M$  to form the final visual context embedding, denoted as the visual local-aggregated context embedding  $\mathbf{v}_g^*$ . Given the text global context embedding  $\mathbf{t}_g^c$  and its enhanced global context embedding  $\mathbf{t}_s^c$ , we perform the fusion operation as follows:

$$\begin{aligned} \mathbf{t}_f^c &= g \cdot \mathbf{t}_s^c + (1 - g) \cdot \mathbf{t}_g^c \\ g &= \text{sigmoid}(\mathbf{W}_g [\mathbf{t}_s^c, \mathbf{t}_g^c] + \mathbf{b}_g) \end{aligned} \quad (9)$$

where  $\mathbf{t}_f^c$  denotes the textual global-fused context embedding obtained by the text global context information  $\mathbf{t}_g^c$  and its enhanced global context information  $\mathbf{t}_s^c$  from visual modality.  $g$  is a gating value to adaptively balance the importance of  $\mathbf{t}_g^c$  and  $\mathbf{t}_s^c$ .  $[\cdot, \cdot]$  means the concatenation operation.

Similarly, as for V-global/T-local contrastive learning, we obtain the textual local-aggregated context embedding  $\mathbf{t}_g^*$  based on  $\{\tilde{\mathbf{t}}_i^*\}_{i=1}^D$ . By fusing  $\mathbf{v}_g^c$  and  $\mathbf{v}_s^c$ , we get the visual global-fused context embedding  $\mathbf{v}_f^c$  similar to Equation 9.

Based on these final global context embeddings, we obtain the contextual-level matching score for a given image-text pair  $(I, T)$ , which is defined as:

$$S_c(I, T) = \frac{1}{2} (\cos(\mathbf{t}_f^c, \mathbf{v}_g^*) + \cos(\mathbf{v}_f^c, \mathbf{t}_g^*)) \quad (10)$$

Then we use the hinge-based triplet ranking loss [10] to enforce the contextual similarity of matched image-text pair to be higher than unmatched ones for contextual alignment:

$$\begin{aligned} \mathcal{L}_{ca} &= \max\left(0, \alpha + S_c(I, \tilde{T}) - S_c(I, T)\right) \\ &\quad + \max\left(0, \alpha + S_c(\tilde{I}, T) - S_c(I, T)\right) \end{aligned} \quad (11)$$

where  $\tilde{T} = \text{argmax}_{d \neq T} s(I, d)$  and  $\tilde{I} = \text{argmax}_{j \neq I} s(j, T)$  are the hardest negatives in a mini-batch for a positive pair  $(I, T)$ , and  $\alpha$  is the margin parameter. Taking the loss objective for shared context enhancement in Equation 8 together, we define the total loss objective in CRA sub-module as:

$$\mathcal{L}_{CRA} = \mathcal{L}_{cs} + \mathcal{L}_{ca} \quad (12)$$

#### D. SelfAlign with Typical Independent-Embedding Models

To prove the effectiveness of our module SelfAlign, we case study on two typical independent-embedding models, the widely compared model VSRN [23] and the state-of-the-art independent-embedding model CAMERA [30].

**VSRN with SelfAlign.** For image encoding, VSRN [23] uses the object embeddings from an object detection model as visual inputs, followed by an FC layer and a graph convolution network with four layers to perform object-relation reasoning. Finally, VSRN utilizes a GRU unit to obtain the visual global embedding. For text encoding, VSRN exploits a word embedding layer to encode word-level embeddings and adopts an LSTM to obtain the contextualized word embeddings. Finally, VSRN utilizes the final hidden state of the LSTM as the textual global embedding. For the loss function, VSRN utilizes the sum of a matching loss on global embeddings via triplet ranking loss and a generation objective from images to texts to jointly align the images and texts. Here, we denote the loss function of VSRN as  $\mathcal{L}_{base}$ . We take the outputs of the FC layer in the image encoder and the outputs of word-embedding layer in the text encoder into LCA sub-module. And we regard the outputs of the last GCN layer and the LSTM as the input of the CRA sub-module.

**CAMERA with SelfAlign.** For the image encoder, CAMERA concatenates the object embeddings and the position embeddings of each object as visual inputs and an FC layer is followed to obtain the visual object embeddings. The word embedding inputs are obtained from the output of pre-trained BERT [8]. Then CAMERA adopts a self-attention layer to perform relation reasoning for image and text respectively. For the loss function, CAMERA adopts the sum of a triplet ranking loss on global embeddings and a diversity regularization loss for cross-modal alignment and the summarization of multi-view descriptions. Here, we also denote the loss function of CAMERA as  $\mathcal{L}_{base}$ . And we take the outputs of the FC layer as LCA sub-module inputs, taking the outputs of the self-attention layer as the inputs of the CRA sub-module.

#### E. Model Training and Inference

The final training loss for baseline models with SelfAlign is defined as:

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_{LCA} + \mathcal{L}_{CRA} \quad (13)$$

where  $\mathcal{L}_{base}$  means the loss objectives of the baseline model.  $\mathcal{L}_{LCA}$  and  $\mathcal{L}_{CRA}$  are defined in Equation 4 and Equation 12.

In the inference stage, the similarity for each image-text pair of baseline with SelfAlign is computed as follows:

$$S(I, T) = S_b(I, T) + S_c(I, T) \quad (14)$$

where  $S_b(I, T)$  is the similarity score evaluated by the baseline model, and  $S_c(I, T)$  denotes our proposed contextual-level similarity score according to Equation 10. It is noteworthy that since our module keeps the advantage of the independent encoding framework of the independent-embedding models, the queried image or text embeddings can pre-computed offline before the inference stage.

#### IV. EXPERIMENT

**Datasets.** We conduct extensive experiments on two benchmark datasets in image-text retrieval: Flickr30K [48] and MS-COCO [24]. Flickr30K consists of 31,783 images collected from the Flickr website and each image is associated with 5 sentences. Following the settings in [10, 21], we utilized 1,000 images for validation, 1,000 images for testing, and the rest for training. MS-COCO contains 123,287 images with 5 captions for each image. Following [10, 21], we take 113,287 images for training, 5,000 images for validation, and 5,000 images for testing, and the results are reported by both averaging over 5 folds of 1K test set images and testing on the full 5K test images as in [10, 21].

**Evaluation Metrics.** To compare with the state-of-the-art models, we adopt the commonly used evaluation metrics in all datasets as [10, 21, 23]. Namely, we adopt Recall at K denoted as R@K to evaluate the performance on both text retrieval (retrieve the most related text given an image query) and image retrieval (retrieve the most related image given a text query) tasks. R@K means the percentage of queries that are correctly matched in the top-K ranking list. We report R@1, R@5, R@10 for all datasets as in [23]. Besides, to comprehensively reveal the overall retrieval performance, we also report another metric R@sum as in [5], defined as the summation of all R@K values in both retrieval tasks.

**Implementation Details.** To perform a fair comparison, for VSRN and CAMERA, we completely preserve their network structures and model settings such as training batch size and other model-related hyper-parameter settings as stated in their original work. We only inject our module into the two baselines as introduced in Section III-D. For both baseline models, the softmax temperature  $\tau_1$  in the LCA sub-module is set to 0.1 as in [4]. The number of concept classes  $K$  is set to 1024, and the number of negative samples  $N$  and the softmax temperature  $\tau_2$  in CRA sub-module are set to 512 and 0.7. All of our experiments are conducted on a single NVIDIA Tesla GPU with 24GB memory and implemented in PyTorch.

##### A. State-of-the-Art Comparison

To verify the effectiveness of SelfAlign, we compare our results with the state-of-the-art models on Flickr30k and MS-COCO in Table I. This table is split into three blocks, from top to bottom, representing independent-embedding models,

interactive-embedding models, and our models (*i.e.* VSRN with SelfAlign and CAMERA with SelfAlign), respectively. Notably, for a fair comparison with the interactive-embedding models, following [23], we achieve our ensemble models by averaging the predicted similarity scores of the two different models obtained by utilizing different seeds for training.

From the comparison between the first block and last block in Table I, we conclude that our module SelfAlign can remarkably improve baseline independent-embedding models on all the metrics, which proves the effectiveness of learning fine-grained correspondences of word-object and global-to-local. Specifically, SelfAlign consistently improves the performance of VSRN and CAMERA by 11.2% and 6.6% in terms of R@sum on Flickr30k, MS-COCO 1K, and MS-COCO 5K. For the strongest baseline model CAMERA, SelfAlign also achieves 9.1%, 4.2%, and 6.6% boosts in terms of R@sum respectively and achieves a new state-of-the-art performance in independent-embedding models. It is worth noting that SelfAlign helps the strongest baseline model CAMERA achieve 1.1%~3.1% boost in terms of R@1 on both retrieval tasks. We also find the performance improvements of R@1 are almost superior than the improvements of R@5/10, indicating that SelfAlign improves the capability of the baseline model to capture fine-grained discrimination on similar images or texts.

From the comparison between the second block and the last block, we conclude that with the help of SelfAlign rather than cross-attention mechanisms, the performance gap between the interactive-embedding models and the independent-embedding models is reduced by a large margin. Specifically, CAMERA with SelfAlign model outperforms the second-best interactive-embedding model ADAPT [44] by 6.7%, 4.8% in terms of R@sum on Flickr30k, MS-COCO 1K. Compared to the-state-of-art interactive-embedding model DIME [31], CAMERA with SelfAlign also achieves comparable performance in terms of R@1 on text retrieval on Flickr30K.

##### B. Efficiency Comparison

To verify the advantage of SelfAlign in keeping the efficiency of independent-embedding models, we construct retrieval latency comparison in the inference phase on baseline models, baseline with SelfAlign models and typical interactive-embedding models. Since the retrieval latency is composed of feature encoding latency and scoring latency, we construct the comparison from these two aspects as shown in Table II and Table III respectively. Since the interactions between the object and word can occur in the feature encoding stage only [31, 43, 44, 49], the scoring stage only [15, 18, 21, 42] or both stages [5, 40], we select representative interactive-embedding models from the above three kinds of methods with comparable accuracy as ours for retrieval latency comparison, *e.g.*, DIME [31], PFAN [42], IMRAM [5].

**Encoding Latency Comparison.** Table II shows the online encoding latency given per query. Besides, Table II also reports the model parameter size which mainly affects the encoding time. Though SelfAlign adds about 36M and 27M parameters and 1~3ms latency to the baselines, VSRN and CAMERA respectively, the encoding latency of the baseline

TABLE I  
COMPARISON WITH EXISTING INDEPENDENT-EMBEDDING MODELS AND INTERACTIVE-EMBEDDING MODELS. OUR RE-IMPLEMENTED INDEPENDENT-EMBEDDING MODELS ARE DENOTED BY THE SUPERSCRIPT ‘\*’. THE HIGHEST RETRIEVAL ACCURACY IN EACH BLOCK IS MARKED WITH UNDERLINE. THE ACCURACY OF OUR MODELS IS MARKED WITH BLUE COLOR WHEN IS BETTER THAN THE BASELINE MODEL.

Model	Flickr30k							MS-COCO 1K							MS-COCO 5K						
	Text Retrieval			Image Retrieval			R@sum	Text Retrieval			Image Retrieval			R@sum	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
<b>Independent-embedding Models</b>																					
Order [38] (2016)	-	-	-	-	-	-	46.7	78.6	88.9	37.9	73.7	85.9	411.7	-	-	-	-	-	-		
2WayNet [9] (2017)	49.8	67.5	-	36.0	55.6	-	208.9	55.8	75.2	0.0	39.7	63.3	0.0	234.0	23.3	50.5	65.0	18.0	43.6		
VSE++ [10] (2018)	52.9	79.1	87.2	39.6	69.6	79.5	407.9	64.6	89.1	95.7	52.0	83.1	92.0	476.5	41.3	69.2	81.2	30.3	59.1		
GXN [39] (2018)	56.8	-	89.6	41.5	-	80.1	-	68.5	-	97.9	56.6	-	94.5	-	42.0	-	84.7	31.7	-		
VSRN [23] (2019)	70.4	89.2	93.7	53.0	77.9	85.7	469.9	74.0	94.3	97.8	60.8	88.4	94.1	509.4	50.3	79.6	87.9	37.9	68.5		
CAMERA [30] (2020)	76.5	95.1	97.2	58.9	84.7	90.2	502.6	75.9	95.5	98.6	62.3	90.9	95.8	519.0	53.1	81.3	89.8	39.0	70.5		
<b>Interactive-embedding Models</b>																					
SCAN_ensemble [21] (2018)	67.4	90.3	95.8	48.6	77.7	85.2	465.0	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3		
CAMP_ensemble [43] (2019)	68.1	89.7	95.2	51.5	77.1	85.3	466.9	72.3	94.8	98.3	58.5	87.9	95.0	506.8	50.1	82.1	89.7	39.0	68.9		
CAAN_ensemble [49] (2020)	70.1	91.6	97.2	52.8	79.0	87.9	478.6	75.5	95.4	98.5	61.3	89.7	95.2	515.6	52.5	83.3	90.9	41.2	70.3		
SGM_ensemble [40] (2020)	71.8	91.7	95.5	53.5	79.6	86.5	478.6	73.4	93.8	97.8	57.5	87.3	94.3	504.1	50.0	79.3	87.9	35.3	64.9		
PFAN_ensemble [42] (2019)	70.0	91.8	95.0	50.4	78.7	86.1	472.1	76.5	96.3	99.0	61.6	89.6	95.2	518.2	-	-	-	-	-		
IMRAM_ensemble [5] (2020)	74.1	93.0	96.6	53.9	79.4	87.2	484.2	76.7	95.6	98.5	61.7	89.1	95.0	516.6	53.7	83.2	91.0	39.6	69.1		
ADAPT_ensemble [44] (2020)	76.6	95.4	97.6	60.7	86.6	92.0	508.9	76.5	95.6	98.9	62.2	90.5	96.0	519.7	-	-	-	-	-		
DIME_ensemble [31] (2021)	81.0	95.9	98.4	63.6	88.1	93.0	520.0	78.8	96.3	98.7	64.8	91.5	96.5	526.6	59.3	85.4	91.9	43.1	73.0		
<b>Ours</b>																					
VSRN*	69.8	88.8	93.6	52.2	78.3	86.1	468.8	73.8	94.1	97.9	60.1	88.1	94.0	508.0	49.7	78.5	87.6	37.2	68.3		
VSRN*+SelfAlign	<b>70.4</b>	<b>91.7</b>	<b>95.9</b>	<b>54.6</b>	<b>81.5</b>	<b>88.4</b>	<b>482.6</b>	<b>74.7</b>	<b>95.1</b>	<b>98.0</b>	<b>62.4</b>	<b>89.4</b>	<b>95.1</b>	<b>514.7</b>	<b>52.4</b>	<b>81.3</b>	<b>89.3</b>	<b>39.6</b>	<b>70.0</b>		
VSRN* Improvement	<b>+0.6</b>	<b>+2.9</b>	<b>+2.3</b>	<b>+2.4</b>	<b>+3.2</b>	<b>+2.3</b>	<b>+13.8</b>	<b>+0.9</b>	<b>+1.0</b>	<b>+0.1</b>	<b>+2.3</b>	<b>+1.3</b>	<b>+1.1</b>	<b>+6.7</b>	<b>+2.7</b>	<b>+2.8</b>	<b>+1.7</b>	<b>+2.4</b>	<b>+1.7</b>		
VSRN*_ensemble	71.0	90.6	94.3	53.9	80.3	87.0	477.1	74.8	95.1	98.3	62.7	89.8	95.0	515.7	51.7	80.8	88.8	39.9	70.4		
(VSRN*+SelfAlign)_ensemble	<b>72.2</b>	<b>92.8</b>	<b>96.6</b>	<b>55.8</b>	<b>82.7</b>	<b>89.0</b>	<b>489.1</b>	<b>75.8</b>	<b>95.5</b>	<b>98.6</b>	<b>64.1</b>	<b>90.5</b>	<b>95.8</b>	<b>520.3</b>	<b>54.3</b>	<b>82.4</b>	<b>90.2</b>	<b>41.3</b>	<b>71.7</b>		
VSRN*_ensemble Improvement	<b>+1.2</b>	<b>+2.2</b>	<b>+2.3</b>	<b>+1.9</b>	<b>+2.4</b>	<b>+2.0</b>	<b>+12.0</b>	<b>+1.0</b>	<b>+0.4</b>	<b>+0.3</b>	<b>+1.4</b>	<b>+0.7</b>	<b>+0.8</b>	<b>+4.6</b>	<b>+2.6</b>	<b>+1.6</b>	<b>+1.4</b>	<b>+1.4</b>	<b>+1.3</b>		
CAMERA*	76.5	93.6	97.3	57.9	84.6	90.5	500.4	74.9	95.4	98.5	62.0	89.9	95.2	515.9	52.4	81.7	89.9	38.8	70.1		
CAMERA*+SelfAlign	<b>79.6</b>	<b>95.1</b>	<b>97.4</b>	<b>59.7</b>	<b>86.2</b>	<b>91.5</b>	<b>509.5</b>	<b>76.8</b>	<b>95.4</b>	<b>98.5</b>	<b>63.1</b>	<b>90.5</b>	<b>95.8</b>	<b>520.1</b>	<b>54.2</b>	<b>82.8</b>	<b>90.6</b>	<b>40.4</b>	<b>71.2</b>		
CAMERA* Improvement	<b>+3.1</b>	<b>+1.5</b>	<b>+0.1</b>	<b>+1.8</b>	<b>+1.6</b>	<b>+1.0</b>	<b>+9.1</b>	<b>+1.9</b>	<b>+0.0</b>	<b>+0.0</b>	<b>+1.1</b>	<b>+0.6</b>	<b>+0.6</b>	<b>+4.2</b>	<b>+1.8</b>	<b>+1.1</b>	<b>+0.7</b>	<b>+1.6</b>	<b>+1.1</b>		
CAMERA*_ensemble	78.3	94.4	97.4	60.5	86.1	91.8	508.4	77.0	96.3	98.6	63.6	90.8	95.8	522.1	55.2	83.1	90.9	40.4	71.4		
(CAMERA*+SelfAlign)_ensemble	<b>81.4</b>	<b>95.6</b>	<b>97.3</b>	<b>61.5</b>	<b>87.1</b>	<b>92.7</b>	<b>515.6</b>	<b>77.7</b>	<b>96.3</b>	<b>98.7</b>	<b>64.3</b>	<b>91.3</b>	<b>96.2</b>	<b>524.5</b>	<b>56.1</b>	<b>83.6</b>	<b>91.0</b>	<b>42.2</b>	<b>72.5</b>		
CAMERA*_ensemble Improvement	<b>+3.1</b>	<b>+1.2</b>	<b>-0.1</b>	<b>+1.0</b>	<b>+1.0</b>	<b>+0.9</b>	<b>+7.2</b>	<b>+0.7</b>	<b>+0.0</b>	<b>+0.1</b>	<b>+0.7</b>	<b>+0.5</b>	<b>+0.4</b>	<b>+2.4</b>	<b>+0.9</b>	<b>+0.5</b>	<b>+0.1</b>	<b>+1.8</b>	<b>+1.1</b>		

TABLE II  
GPU TIME IN EARLY FEATURE ENCODING STAGE PER QUERY ON FLICKR30K TEST SET.

Model	Param.(M)	# of candidates			
		1K(ms)	10K(ms)	100K(ms)	1000K(ms)
PFAN [42]	12.8	2.5	2.5	2.5	2.5
IMRAM [5]	21.8	$4.2 \times 10^3$	$4.2 \times 10^4$	$4.2 \times 10^5$	$4.2 \times 10^6$
DIME [31]	116.3	$3.0 \times 10^3$	$3.0 \times 10^4$	$3.0 \times 10^5$	$3.0 \times 10^6$
VSRN [23]	137.7	7.3	7.3	7.3	7.3
VSRN+SelfAlign (ours)	173.6	8.7	8.7	8.7	8.7
CAMERA [30]	156.2	25.7	25.7	25.7	25.7
CAMERA+SelfAlign (ours)	183.5	28.2	28.2	28.2	28.2

TABLE III  
TIME IN SCORING PER QUERY ON 100K CANDIDATES. INTER. DENOTES THE NUMBER OF THE INTERACTIONS BETWEEN WORDS AND REGIONS. FLOPS DENOTES THE NUMBER OF FLOATING-POINT OPERATIONS. DIM. DENOTES THE DIMENSION OF THE SIMILARITY CALCULATION VECTOR.

Model	Inter.	FLOPs	Dim.	GPU TIME(ms)
PFAN [42]	$32 \times 36$	$\times 1152$	1024	$3.3 \times 10^3$
IMRAM [5]	$32 \times 36 \times 3$	$\times 3456$	1024	$9.9 \times 10^3$
DIME [31]	$1 \times 1$	$\times 1$	256	1.0
VSRN [23]	$1 \times 1$	$\times 1$	2048	2.0
VSRN+SelfAlign (ours)	$1 \times 1$	$\times 1$	6144	2.3
CAMERA [30]	$1 \times 1$	$\times 1$	2048	2.0
CAMERA+SelfAlign (ours)	$1 \times 1$	$\times 1$	6144	2.3

with SelfAlign models still keeps invariant to the number of candidates, which is much lower than the accuracy comparable interactive-embedding models. This is because SelfAlign preserves the independent feature encoding architecture and does not add any cross-modal interactions in the encoding stage. Therefore, the query embedding can be encoded independently without the interactions with the queried candidates, and the embeddings of the queried candidates can be pre-computed offline without the consuming of online encoding latency. In contrast, those models performing cross-modal interactions

during encoding stage, like IMRAM and DIME, need to encode each text-image pair, resulting in the encoding latency is linearly related to the number of candidates.

**Scoring Latency Comparison.** Table III shows the scoring latency of per query on 100K candidates. The scoring time is linear with the number of interactions between image embeddings and text embeddings. The results show that SelfAlign increases about 0.3ms latency to baseline models while both baseline models with SelfAlign are still 1000 times faster than the model PFAN and IMRAM, which have cross-modal token-wise interactions during the scoring stage. This is because SelfAlign still keeps the baseline model to perform the computation of similarity scores by using simple similarity calculation operations like dot product.

### C. Accuracy and Efficiency Joint Comparison

To prove the good balance between accuracy and efficiency for image-text retrieval of our module, we visualize the accuracy and efficiency jointly on Flickr30K and MS-COCO 5K in Figure 3. Specifically, the trade-off models include VisualSparta [28], ISERI\_inflate [26], ISERI\_fast [26], LightDOT [34]. VisualSparta [28], ISERI\_inflate [26] are late-interaction trade-off models, where there are no cross-modal token-wise interactions in the feature encoding stage but in the scoring stage. LightDOT [34] is a pre-trained re-ranking based method, which utilizes a pre-trained independent-embedding model to coarsely rank in the first stage and then utilizes another pre-trained interactive-embedding model UNITER [7] to finely rank in the second stage. ISERI\_fast [26] is also an independent-embedding model but is obtained by performing knowledge distillation twice under the extra model supervision of ISERI\_inflate [26]. This enables ISERI\_fast to obtain more

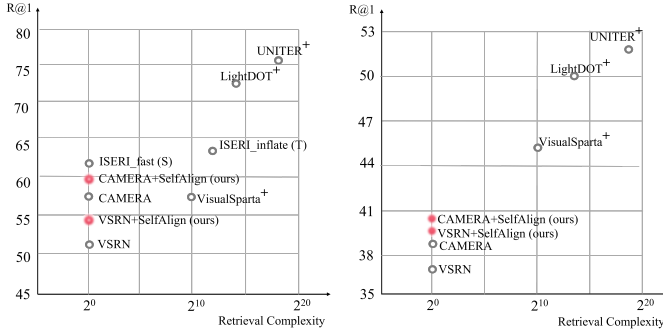


Fig. 3. Image retrieval results of Recall@1 of (a) Flickr30K and (b) MSCOCO 5K. The retrieval complexity refers to the number of cross-modal interactions included in the early feature encoding stage and late scoring stage. The superscript ‘+’ denotes pre-trained models on huge-scale datasets. ‘S’ and ‘T’ represent the student model and the teacher model respectively.

TABLE IV  
ABLATION STUDY OF CAMERA WITH SELFALGIN ON FLICKR30K.

Model	Param. (M)	GPU TIME (ms)	Text Retrieval			Image Retrieval			R@sum
			R@1	R@5	R@10	R@1	R@5	R@10	
#0 Full Model	183.5	2.3	79.6	95.1	97.4	59.7	86.2	91.5	509.5
Ablation of Sub-Module									
#1 w/o LCA	183.5	2.3	76.9	94.5	97.5	59.6	84.8	90.9	504.2
#2 w/o CRA	156.2	2.0	77.0	95.1	97.2	58.7	84.6	90.6	503.2
Ablation of LCA									
#3 w/ Concept O2W	183.5	2.3	77.7	94.3	97.4	60.5	85.3	91.5	506.8
#4 w/ Concept Dual	183.5	2.3	78.1	94.6	97.3	60.4	85.4	91.3	507.1
#5 w/ noun+adj+verb	183.5	2.3	77.2	94.2	97.2	59.2	85.5	91.2	504.5
#6 w/ noun	183.5	2.3	76.4	94.8	97.0	59.2	84.9	91.0	503.4
Ablation of CRA									
#7 w/o $\mathcal{L}_{cs}$	156.2	2.2	76.1	94.0	97.9	59.4	85.3	90.9	503.7
#8 w/o $\mathcal{L}_{ca}$	183.5	2.3	76.1	94.6	97.3	58.5	84.1	89.8	500.5
#9 w/o $\mathcal{L}_{tg}$ Atd Align	169.9	2.2	77.4	95.1	97.4	59.4	84.7	90.8	504.8
#10 w/o $\mathcal{L}_{vg}$ Atd Align	169.9	2.2	78.2	94.6	97.4	59.3	84.8	90.8	505.1
#11 CAMERA*	156.2	2.0	77.1	93.5	96.3	58.6	84.4	90.6	500.5

accurate fine-grained information under the supervision of the teacher model ISERI\_inflate compared with our models.

In Figure 3, the  $x$ -axis represents the retrieval complexity referring to the average number of cross-modal interactions per candidate when given a query to retrieve the ground-truth from 1000 candidates. The number of cross-modal interactions is the sum of the number of interactions in both of the feature encoding stage and the scoring stage. The  $y$ -axis denotes the results of R@1 in the image retrieval task. The red dots represent our models. As shown in Figure 3, SelfAlign enables both baseline models improving the accuracy without increasing retrieval complexity while other trade-off models sacrifice retrieval efficiency for cross-modal token-wise interactions. SelfAlign explores the cross-modal semantic correspondences and improves the retrieval efficiency by achieving fine-grained alignment matching fine-grained correspondences in the early feature encoding stage.

#### D. Ablation Study

We conduct ablation study to evaluate the effectiveness of essential components in SelfAlign. The results on Flickr30k are shown in Table IV. We use CAMERA+SelfAlign as the full model for all the following variants:

- **w/o LCA (model ‘#1’)**: this model removes the Local Concept Alignment sub-module in SelfAlign.
- **w/o CRA (model ‘#2’)**: this model removes the Contextual Relation Alignment sub-module in SelfAlign.
- **w/ Concept O2W (model ‘#3’)**: this model changes the direction of concept alignment learning in LCA sub-

module from word-object to object-word. Specifically, we choose the most similar word as the correspondence for each object and perform clustering on word embeddings for fine-grained alignment learning.

- **w/ Concept Dual (model ‘#4’)**: this model preserves these two directions of concept alignment learning.
- **w/ noun+adj+verb (model ‘#5’)**: this model performs concept alignment learning only with nouns, verbs and adjectives instead of all the words.
- **w/ noun (model ‘#6’)**: this model performs concept alignment learning only with nouns.
- **w/o  $\mathcal{L}_{cs}$  (model ‘#7’)**: this model removes the global-to-local contrastive loss  $\mathcal{L}_{cs}$  in Equation 8 for shared context enhancement in CRA sub-module.
- **w/o  $\mathcal{L}_{ca}$  (model ‘#8’)**: this model removes that the contextual alignment loss  $\mathcal{L}_{ca}$  in Equation 11 for shared context alignment learning in CRA sub-module.
- **w/o  $\mathcal{L}_{tg}$  Atd Align (model ‘#9’)**: this model removes T-global/V-local contrastive learning attended context alignment. There is no projection for  $t_g^c$  in Equation 5,  $\mathcal{L}_{tg}$  in Equation 6, text fusion in Equation 9 and the similarity computation in the first term of Equation 10.
- **w/o  $\mathcal{L}_{vg}$  Atd Align (model ‘#10’)**: this model removes V-global/T-local contrastive learning attended context alignment, which is symmetrical with model ‘#9’.

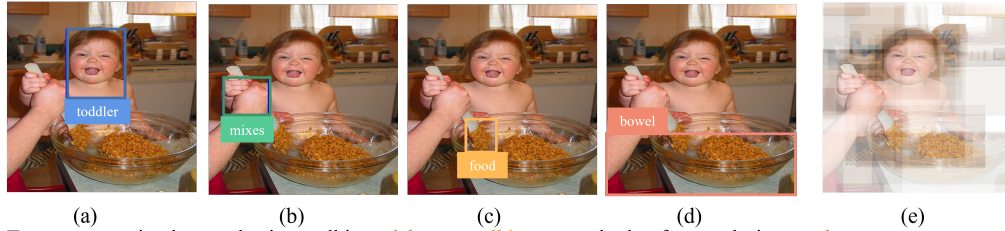
Models in the first block are designed to evaluate the contribution of each sub-module in SelfAlign. We observe that the R@sum value of models ‘#1-#2’ all significantly decrease by 5.3% and 6.3%, but they still outperform the baseline model ‘#11’. It shows that both alignment sub-modules are effective for baseline to extract different levels of fine-grained correspondences and the combination of them can further improve retrieval accuracy. It proves the effectiveness and complementarity of concept level and contextual level alignment information learned in LCA and CRA sub-module.

Models in the second block evaluate the influence of the single direction and the effectiveness of utilizing all words to perform word-object alignment learning in LCA. The performance of either model ‘#3’ or model ‘#4’ decreases slightly by 2.7% and 2.4% respectively, which demonstrates word-to-object alignment captures more accurate concept alignment information. Besides, The performance of either model ‘#5’ or model ‘#6’ decreases by 5.0% and 6.1%, which demonstrates the effectiveness of making the most use of text information via utilizing all words to align regions in LCA sub-module.

Models in the third block evaluates the influence of the key components in the CRA sub-module. The performance of model ‘#7’ and model ‘#8’ decreases remarkably compared with the full model. The results show that both shared context enhancement by the global-to-local contrastive loss and shared context alignment learning are essential for contextual alignment. The performance of model ‘#8’ is decreased to the results as the baseline model ‘#11’ in terms of R@sum, which indicates that CRA sub-module only with global-to-local contrastive learning objective  $\mathcal{L}_{cs}$  is not only ineffective in context-level alignment but also damaging to the concept-level alignment in the LCA sub-module. The reason is that only with  $\mathcal{L}_{cs}$  in the model ‘#8’ takes the local contextual



Q1: A toddler mixes some food in a bowl.



Q2: Two men wearing hats and using walking sticks are walking near a body of water during sundown.



Fig. 4. Visualization of similarity in LCA and CRA alignment module. (a) (b) (c) (d) visualize the most matched word-to-region pair results in LCA, where each text concept is used as a query to find the most similar image region and the region outlined in the same color. (e) visualizes the global to local relevance of CRA at the training stage, where the region brightness represents the similarity strength.

embeddings from unmatched image-text pairs as the negative samples according to Equation  $L_{cs}$  to learn the shared context information in the image-text pair, which ignores that these local contextual embeddings could mainly include the correct concept information in the concept alignment of the LCA sub-module. This conflict makes both the LCA and CRA sub-module ineffective. Compared to model ‘#2’, the performance of model ‘#9’ and model ‘#10’ improves slightly even when the models only preserve half part of CRA sub-module. These results verify the effectiveness of preserving the joint learning mode of global-to-local contrastive loss. Moreover, compared to the full model, model ‘#9’ and model ‘#10’ lead to the performance degradation, which demonstrates that these two context-level alignment are effective and capture complementary contextual information from vision global supervision and text global supervision.

Moreover, we observe that the LCA sub-module does not take extra parameters for the baseline model as it only involves in  $\mathcal{L}_{LCA}$  defined in Equation 4 for local concept alignment learning. The CRA sub-module causes the increments of parameters and retrieval latency for baseline model CAMERA by 27.3M and 0.3ms, respectively, which is due to the feature linear projections defined in Equation 5 and feature fusion based on the gate mechanism defined in Equation 9.

### E. Alignment Quality Analysis

#### Alignment visualization of the concept representation.

We utilize t-SNE [36] to visualize region and word embeddings of some frequent concepts for intuitively analyzing the concept-level alignment, as shown in Figure 5. The embeddings are obtained by CAMERA model and CAMERA+SelfAlign model. We conclude that SelfAlign pulls the distance of same semantic concepts of different modalities. For example, in Figure 5 (b), the concept ‘water’, ‘dirt’, and ‘shorts’ are densely clustered by CAMERA+SelfAlign while CAMERA can not.

#### Alignment visualization in LCA and CRA sub-module.

The detailed alignment process is interpretable by visualizing

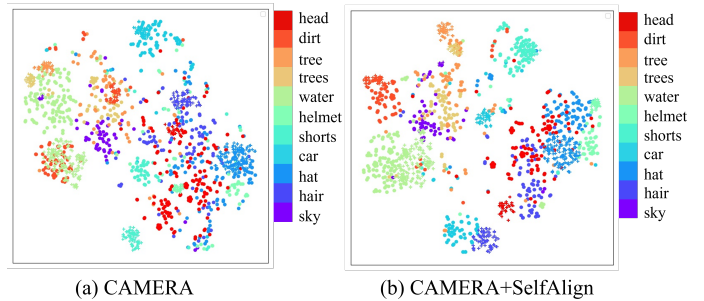


Fig. 5. T-SNE visualization of concept representations. ‘Cross’ means textual concept, and ‘circle’ means visual concept.

the conceptual and contextual similarity scores at LCA sub-module and CRA sub-module. The results are shown in Figure 4. Specifically, given an image-text pair in the inference phase, the most similar region for each word is identified in LCA sub-module. The global to local similarity map is computed in CRA sub-module, which supports the explicit visualization and reveals the alignment process. From the first four columns, we observe that the words align to the proper regions with the highest similarity. For example, in the first example, our model finds appropriately matched regions on nouns like ‘toddler’, ‘food’, ‘bowl’, as well as the action word like ‘mixed’. In the CRA sub-module, we observe that semantically matched image regions align well to the text global context embedding, while irrelevant image regions are suppressed. These examples prove that SelfAlign learns the concept-level word-object correspondences in LCA sub-module and composites the essential regions to understand the global context information in CRA.

**Qualitative retrieval results analysis.** The qualitative results from text-to-image retrieval and the image-to-text retrieval on Flickr30K are illustrated in Figure 6 and Figure 7, respectively. From both retrieved results, it’s clear that CAMERA+SelfAlign retrieves the correct candidates to a more forward position. Moreover, for those candidates with similar

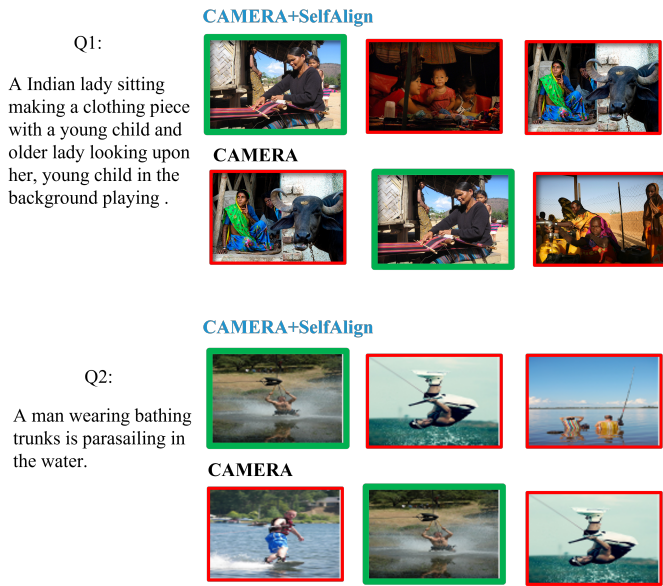


Fig. 6. Qualitative comparison of text-to-image retrieval between the baseline model CAMERA and CAMERA+SelfAlign on Flickr30K. We show the top-3 retrieved images for each text query. The truly matched results are marked in green boxes and the falsely matched results are in red boxes.

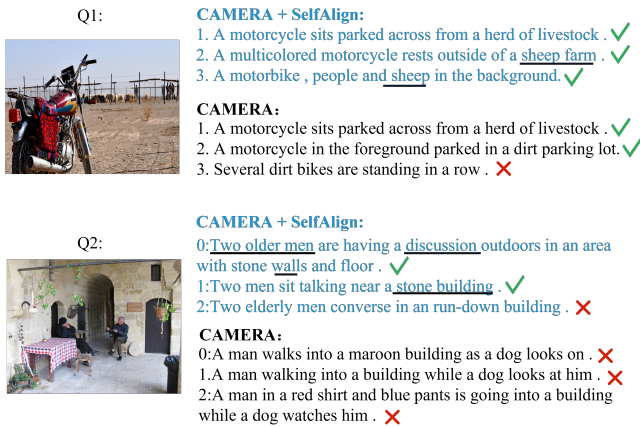


Fig. 7. Qualitative comparison of image-to-text retrieval between the baseline model CAMERA and CAMERA+SelfAlign on Flickr30K. We show the top-3 ranked texts for each image query. The truly matched sentences are marked with checks and the falsely matched results are with cross. The concepts in the retrieved text of CAMERA+SelfAlign that differ from the baseline model are marked with underline.

scenes, SelfAlign enables the baseline model to distinguish the fine-grained discrimination among the candidates well, which verifies that SelfAlign is effective in fine-grained cross-modal information retrieval.

## V. CONCLUSION

In this paper, we propose a fine-grained image-text alignment module SelfAlign for fast and accurate image-text retrieval. We design two collaborative sub-modules to learn complementary alignment information from both conceptual and contextual level in a self-supervised manner, which improves the retrieval accuracy while keeps the retrieval efficiency. SelfAlign is model-agnostic and generic to incorporate

with various independent-embedding retrieval approaches. Our module consistently boosts the accuracy of the strongest non-pre-training independent-embedding model. How to extend SelfAlign on other cross-modal retrieval tasks such as video-text retrieval will be our future work.

## REFERENCES

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.
- [2] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, pages 15509–15519, 2019.
- [3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 139–156, 2018.
- [4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, pages 9912–9924, 2020.
- [5] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*, pages 12655–12663, 2020.
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.
- [7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.
- [9] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, pages 4601–4611, 2017.
- [10] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, page 12, 2018.
- [11] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomás Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013.
- [12] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulic, and Iryna Gurevych. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *TACL*, pages 503–521, 2022.
- [13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020.
- [14] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.
- [15] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, pages 2310–2318, 2017.
- [16] Zhong Ji, Kexin Chen, and Haoran Wang. Step-wise hierarchical alignment network for image-text matching. In *IJCAI*, pages 765–771, 2021.
- [17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021.
- [18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.
- [19] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.
- [20] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*, pages 4392–4412, 2020.
- [21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.
- [22] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. In *ICLR*, 2022.

- [23] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4654–4662, 2019.
- [24] 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 *ECCV*, pages 740–755, 2014.
- [25] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10921–10930, 2020.
- [26] Haoliang Liu, Tan Yu, and Ping Li. Inflate and shrink: Enriching and reducing interactions for fast text-image retrieval. In *EMNLP*, pages 9796–9809, 2021.
- [27] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*, pages 1950–1959, 2019.
- [28] Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. Visualsparta: An embarrassingly simple approach to large-scale text-to-image search with weighted bag-of-words. In *ACL/IJCNLP*, pages 5020–5029, 2021.
- [29] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, pages 9826–9836, 2021.
- [30] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. Context-aware multi-view summarization network for image-text matching. In *ACM MM*, pages 1047–1055, 2020.
- [31] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *SIGIR*, pages 1104–1113, 2021.
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.
- [33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *TPAMI*, pages 1137–1149, 2016.
- [34] Siqu Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *NAACL*, pages 982–997, 2021.
- [35] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *ICML*, pages 10268–10278, 2021.
- [36] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *J Mach Learn Res*, pages 2579–2605, 2008.
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.
- [38] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.
- [39] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACM MM*, pages 154–162, 2017.
- [40] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, pages 1508–1517, 2020.
- [41] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021.
- [42] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *IJCAI*, pages 3792–3798, 2019.
- [43] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, pages 5764–5773, 2019.
- [44] Jonas Wehrmann, Camila Kolling, and Rodrigo C Barros. Adaptive cross-modal embeddings for image-text alignment. In *AAAI*, pages 12313–12320, 2020.
- [45] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *ACM MM*, pages 2088–2096, 2019.
- [46] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.
- [47] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, pages 6210–6219, 2019.
- [48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, pages 67–78, 2014.
- [49] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *CVPR*, pages 3536–3545, 2020.



**Jiamin Zhuang** is currently studying for a Ph.D. degree in the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Jiamin Zhuang received her B.S. degree in network engineering from Henan University, China, in 2016. Her research interests mainly focus on cross-modal retrieval.



**Jing Yu** is currently an associate professor in the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Jing Yu received her B.S. degree in Automation Science from Minzu University, China, in 2011, and got her M.S. degree in Pattern Recognition from Beihang University, China in 2014. She received her Ph.D. degree in the University of Chinese Academy of Sciences, China, in 2019. Her research interests mainly focus on cross-modal understanding, including visual question answering, cross-modal information

retrieval, scene graph generation, etc.



**Yang Ding** is currently studying for a master's degree in the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Yang Ding received his B.S. degree in chemistry from Wuhan University, China, in 2016. His research interests mainly focus on knowledge-based visual question answering.



**Xiangyan Qu** is currently studying for a Ph.D. degree in the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Xiangyan Qu received his B.S. degree in Internet of Things Professional from University of Science and Technology Beijing, China, in 2017. His research interests mainly focus on scene recognition.



**Yue Hu** is a Professor in the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Her research interests are in the area of natural language processing and social network analysis.