Decoupled Cross-modal Phrase-Attention Network for Image-Sentence Matching

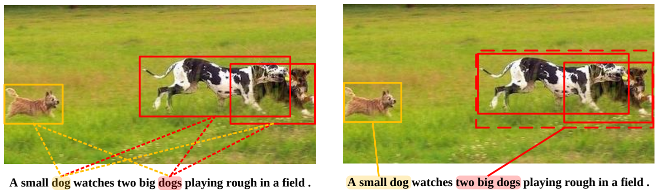
Zhangxiang Shi, Tianzhu Zhang, Xi Wei, Feng Wu, and Yongdong Zhang

***Abstract*—The mainstream of image and sentence matching studies currently focuses on fine-grained alignment of image regions and sentence words. However, these methods miss a crucial fact: the correspondence between images and sentences does not simply come from alignments between individual regions and words but from alignments between the phrases they form respectively. In this work, we propose a novel Decoupled Cross-modal Phrase-Attention network (DCPA) for image-sentence matching by modeling the relationships between textual phrases and visual phrases. Furthermore, we design a novel decoupled manner for training and inferencing, which is able to release the trade-off for bi-directional retrieval, where image-to-sentence matching is executed in textual semantic space and sentence-to-image matching is executed in visual semantic space. Extensive experimental results on Flickr30K and MS-COCO demonstrate that the proposed method outperforms state-of-the-art methods by a large margin, and can compete with some methods introducing external knowledge.**

I. Introduction

With the spring up of multimedia data from social media and web applications in recent years, image-sentence matching has become an emerging task which tries to accurately measure correspondences between image and sentence pairs [1]–[3]. It is one of the fundamental problems for multimedia content understanding, and can provide inspiration for tasks such as cross-modal retrieval [4]–[7], visual captioning [8], [9], visual grounding [10], [11] and visual question answering [12]–[14]. Although significant progress has been achieved in this task, further studies are still urged, as accurately measuring the similarity score of the given image and sentence pair requires the understanding of visual, textual semantics and their relationships.

One of the most challenging difficulties for this task is so-called "heterogeneity gap" which is caused by the different data form of images and sentences, making it impossible to measure the similarities between images and sentences directly [15]. About this problem, a common solution is to seek a joint semantic space in which the image-sentence pairs sharing similar semantics are closer than those mismatched [4], [16]. Following [2], we classify these methods as "one-to-one" and "many-to-many" methods. One-to-one methods learn global representations for image and sentence, then align them using the pairwise similarity information. Particularly, they usually adopt pre-trained neural networks like CNN and RNN to encode the whole image or sentence into a common feature space. However, independently embedding heterogeneous instances might lead to suboptimal results. Besides, these works only coarsely capture the alignment between the whole image and sentence and thus ignore the fine-grained interactions between image regions and sentence words.


A small dog watches two big dogs playing rough in a field .
A small dog watches two big dogs playing rough in a field .

(a) Region-Word Correspondences (b) Phrase-Phrase Correspondences

Fig. 1: (a) Region-Word correspondences might be inappropriate and lead to ambiguity. (b) Phrase-Phrase correspondences can be more precise covering region-word correspondences.

To overcome the above issues, many-to-many methods have been proposed by learning local region-word relationships [1], [17]–[20]. Generally, these methods can be roughly grouped into inter-modal based methods, intra-modal based methods, and methods that consider both of them. Inter-modal based methods consider inter-modal correspondences among visual fragments and textual fragments [1], [2], [21], [22], while intra-modal based methods take advantage of intra-modal correspondences [3], [23]–[25]. More methods take both inter-modal and intra-modal correspondences into consideration [13], [15], [26]–[31]. In addition, some approaches attempt to introduce external knowledge including additional dataset used to train networks to capture image salience information [32] or an off-the-shelf tool for sentence parsing to build better representations for image and sentence [33], while some approaches resort to pre-training tasks [34]–[41]. Thanks to the effective backbones for extracting fragments of images and sentences, they have achieved considerable success and have become mainstream. However, these methods [1], [2], [21] ignore one thing: aligning isolated regions with isolated words directly might be inappropriate and lead to ambiguity. As shown Figure 1 (a), there is no specific information about which bounding box should the word "dog" be aligned to as each bounding box shares the same semantics "dog". It is even worse for the word "dogs" as it should be aligned to two bounding boxes rather than one. This happens again when we attempt to find corresponding words for each bounding box.

Manuscript received *** **, 202X; revised *** **, 202X; accepted *** **, 202X.

Zhangxiang Shi, Tianzhu Zhang, Xi Wei, Feng Wu, and Yongdong Zhang are with school of information science, University of Science and Technology of China, Hefei 230027, China (e-mail: szxszx@mail.ustc.edu.cn; tzzhang@ustc.edu.cn; wx33921@mail.ustc.edu.cn; fengwu@ustc.edu.cn; zhyd73@ustc.edu.cn).

Corresponding author: Yongdong Zhang.

On the contrary, we consider the alignment between visual phrases and textual phrases. As shown in Figure 1 (b), we aggregate regions into visual phrases and words into textual phrases, and the inter-modal correspondences can be more precise. In this case, the textual phrase "two big dogs" is aligned to the visual phrase covering two red bounding boxes correctly while the textual phrase "a small dog" is also correctly aligned to the visual phrase covering one yellow bounding box. Note that both visual phrases and textual phrases can also contain only one element. To sum up, we can see that the fragments of sentences and images may not tell the correspondences between sentences and images accurately. Instead, we should aggregate the semantically related words in the sentence and the semantically related regions in the image which we call them "**textual phrases**" and "**visual phrases**" respectively for convenience, and find the correspondences between them. Furthermore, existing methods [1], [3], [13], [24], [26], [42] perform semantic alignment in a joint embedding space, which may lead to suboptimal results due to the heterogeneity gap and the trade-off between I2T (Image to Text retrieval) and T2I (Text to Image retrieval) subtasks.

Motivated by the above discussions, we propose a novel Decoupled Cross-modal Phrase-Attention network for image and sentence matching. Our network can model the relationships between textual phrases and visual phrases, and refine their representations based on data from another modality in a decoupled manner. To achieve this goal, we design two effective phrase-attention modules, including an intra-modal phrase-attention module and an inter-modal phrase-attention module. Specifically, we first extract features of salient image regions and sentence words. In the intra-modal phrase-attention module, we use an attention mechanism to capture textual phrases in sentences and visual phrases in images separately and then align them using pairwise similarities. While in the inter-modal phrase-attention module, we design two decoupled branches for the I2T matching subtask and the T2I matching subtask separately. For the I2T branch, we first use an attention mechanism to aggregate image regions into visual phrases, then for each visual phrase, we make it attend to all the words from a sentence, which allows each visual phrase to find its own best-matched textual phrase in this sentence. After that, we map these visual phrases into the textual semantic space, utilizing their own corresponding best-matched textual phrases. Finally, we calculate similarities with respect to each pair of visual phrases and textual phrases, and the global image-to-sentence similarity is aggregated from these local similarities. The T2I branch is handled similarly and will not be covered here. Note that, unlike existing works [1], [13], [43], we perform the I2T and T2I subtasks in an asymmetric way with different modules, achieving them in the textual guided embedding space and the visual guided embedding space respectively, which not only reduces the difficulty of optimization but also releases the trade-off between I2T and T2I subtasks and results in better performance. We refer to this property as "**decoupled** manner" for simplicity.

Our contributions can be summarized into three-folds. (1) We propose a novel decoupled cross-modal phrase-attention network for image-sentence matching by jointly modeling the intra-modal phrase-attention and the inter-modal phrase-attention in a unified deep model. (2) To the best of our knowledge, this is the first work that is able to not only construct visual phrases and textual phrases and find their correspondences utilizing multi-modal attention mechanisms, but also align visual and textual phrases precisely in a decoupled manner. (3) Extensive experimental results on Flickr30K and MS-COCO demonstrate that our proposed method outperforms state-of-the-art methods by a large margin, and can compete with some methods introducing external knowledge.

II. Related Work

In this section, we briefly review the methods related to ours including one-to-one matching methods, many-to-many matching methods, and pre-training methods.

One-to-One Matching Methods. Plentiful early work extracts global features for image and sentence, and then attempt to associate them using supervision information in the form of pairwise constraints [44]–[47]. For instance, Frome et al. [48] use CNN and Skip-Gram [49] to extract image and sentence representations and then associate them utilizing projection layers with a structured objective. While Yan et al. [47] associate features of image and sentence with deep canonical correlation analysis where the matched image-sentence pairs have a high correlation. For similar purposes, Klein et al. [45] use Fisher Vectors (FV) to obtain a discriminative sentence representation. Faghri et al. [43] attempt to use hard negative mining in the bidirectional triplet ranking loss and achieve a significant improvement. While in [50] and [46], GAN [51] is combined with multimodal embedding learning to learn more discriminative representations for image and sentence. Recently, Chun et al. [52] propose to use probabilistic intra-modal global embedding to capture one-to-many correspondences in images and their captions. However, these methods ignore the fact that the global similarity between an image and a sentence arises from the aggregation of latent vision-language correspondences between image regions and sentence words.

Many-to-Many Matching Methods. It has become mainstream to consider the fine-grained correspondences between image regions and sentence words. Many researchers consider inter-modal correspondence among visual fragments and textual fragments [1], [2], [21], [22]. Huang et al. [2] present a multi-modal LSTM with context-modulated attention scheme to attend to salient regions and words sequentially. While SCAN [1] uses a stacked cross attention mechanism to discover all the alignments between salient regions and words in one step. But they ignores the interactions within image regions or sentence words. Some other researchers take advantage of intra-modal correspondence [3], [23]–[25], [53], [54]. For instance, Wu et al. [24] constructs two independent branches to capture the attention between image regions and the attention between sentence words respectively. Li et al. [25] first perform local reasoning between image regions and then use GRU to perform global semantic reasoning to capture both key objects and global semantic concepts of a scene. Chen et al. [53] perform generalized pooling operations within each modality fragments to discover effective

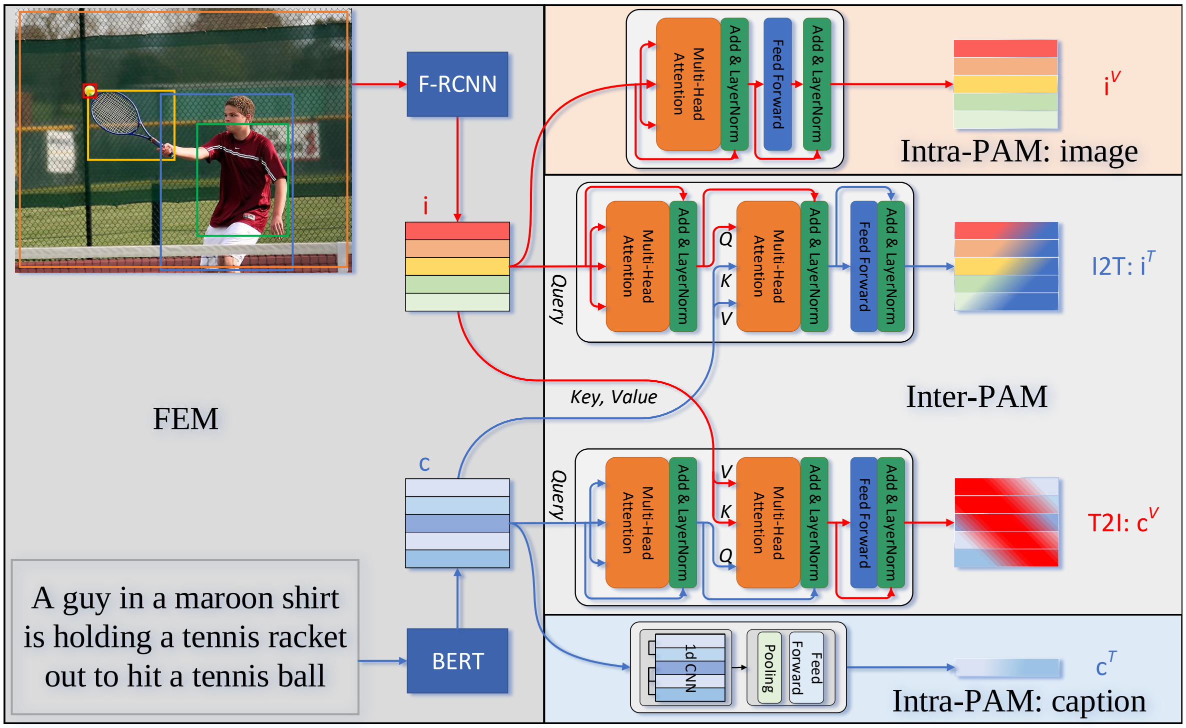


Fig. 2: The framework of our proposed Decoupled Cross-modal Phrase-Attention network. The whole framework can be divided into three parts, including a feature extraction module, an intra-modal phrase-attention module, which is designed separately for image branch and text branch, and an inter-modal phrase-attention module working in a decoupled manner.

image and sentence global embeddings, while Yan et al. [54] propose a novel attention scheme which can optimises the intra-modal attention weights directly towards the evaluation metrics. These methods ignore the interactions between image regions and sentence words. More methods consider both inter-modal and intra-modal correspondence [13], [15], [26]–[31], [42], [55], [56]. Among them, MIA [27], LXMERT [15], MMCA [26], MCAN [13], CAAN [30] achieve inter-modal interaction and intra-modal interaction by combining inter-attention and intra-attention jointly. While Li et al. [29], MLI [28], M^2 Transformer [31] explore higher-order attention, considering the inherent structure of images and sentences. In addition, IMRAM [42] executes an attention refinement in an iterative manner to simulate the complex process of human attention. WCGL [55] and SGRAF [56] construct a graph to model relationships in both textual and visual modalities, performing similarity reasoning on the graph. Besides, some approaches attempt to introduce external knowledge to build better representation for image and sentence [32], [33], [57]. GSMN [33] and SMFEA [57] uses an off-the-shelf Stanford CoreNLP [58] to guide the syntactic structure constructing and matching of sentences and images. SAN [32] trains a saliency network with an external MSRA10K dataset [59] to provide the visual saliency information. However, none of these approaches matches image and sentence by constructing and aligning visual phrases and textual phrases. Also, existing methods perform semantic alignment in the joint embedding space, without using dedicated I2T and T2I matching branches. Compared with them, our approach implements the construction and alignment of visual phrases and textual phrases in a decoupled manner, without any external knowledge or datasets.

Pre-training Methods. Recently, some researchers have tried pre-training methods for visual and language cross-modal tasks [34]–[41], [60]–[63]. They generally rely on a multi-layer transformer, pretrained with large-scale multi-modal datasets and then finetuned with various kinds of downstream tasks. Although these methods have been enormously successful, they come at a huge computational and memory cost and rely heavily on extra large datasets. Again, they do not construct and align visual phrases and textual phrases and do not achieve I2T and T2I tasks in a decoupled manner.

III. Our Proposed Approach

As shown in Figure 2, our Decoupled Cross-modal Phrase-Attention network consists of three modules including Feature Extraction Module (FEM), Intra-Modal Phrase-Attention Module (Intra-PAM), and Inter-Modal Phrase-Attention Module (Inter-PAM). The Intra-PAM aggregates regions into visual phrases and words into textual phrases and aligns them with pairwise similarity constraints, and the Inter-PAM implements the construction and interaction for multi-modal phrases to achieve cross-modal alignment. We describe these modules and our alignment objective and training strategy as follows.

A. Feature Extraction Module

Given an image and sentence pair, we first extract their fine-grained representations. For an image I, we use the bottom-up attention model [12] pre-trained on Visual Genome [64] to extract region features. The output can be represented as $\mathbf{O} = \{o_1; o_2; \cdots; o_m\}$, where o_k is the feature for the k-th region. A fully-connect layer is added to reduce their dimensions and the output is denoted as $i = \{r_1; r_2; \cdots; r_m\}$. We use the pretrained BERT [65] model to get a fine-grained representation of the sentence. To be more specific, we first use Word-Piece tokens [66] of the sentence T as its fragmented token embedding. The embedding for each word in the sentence is the combination of its token embedding, position embedding, and segment embedding, following the setting of BERT. Then the pretrained BERT is used to refine representations of words in this sentence, taking full advantage of contextual information. The output is denoted as $c = \{w_1; w_2; \cdots; w_n\}$.

B. Intra-Modal Phrase-Attention Module

In this section, we introduce how we design different phrase-attention modules for images and for sentences, respectively. These two branches both employ attention mechanism to construct modal-specific phrases, but with different implementations.

Visual Phrase Construction. In this branch, each region in an image firstly attends to all the regions from the same image and then aggregates them to form a visual phrase for itself. As a consequence, the output of the Intra-PAM for an image is a set of visual phrases with respect to every region in it. Specifically, we use an attention module to implement this. The attention mechanism may combine existing fragments into contextual fragments, usually in a linear weighting way, and is in conjunction with other nonlinear transformations. One way to achieve this is to first map original fragments to queries and key-value pairs, then use a weighted sum of the values as outputs, where the weighting factors depend on the interaction between queries and keys. Specifically, we use the transformer encoder [67] to construct visual phrases, and briefly review its structure below. As shown in the upper right part of Figure 2, the transformer encoder consists of a multi-head attention sublayer and a position wise feed-forward sublayer. The multi-head attention sublayer is the key module to implement the attention mechanism. Given a set of region features $i = \{r_1; r_2; \cdots; r_m\}$, where $r_k \in \mathbf{R}^{1\times d_f}$ and $i \in \mathbf{R}^{m\times d_f}$, the functions of this sublayer can be explained by following formulas:

$$\mathbf{Q}_j = i\mathbf{W}_j^Q, \mathbf{K}_j = i\mathbf{W}_j^K, \mathbf{V}_j = i\mathbf{W}_j^V, \quad (1)$$

$$\text{Attention}\left(\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j\right) = \text{softmax}\left(\frac{\mathbf{Q}_j\mathbf{K}_j^T}{\sqrt{d_k}}\right)\mathbf{V}_j. \quad (2)$$

Where $\mathbf{W}_j^Q \in \mathbf{R}^{d_f\times d_q}$, $\mathbf{W}_j^K \in \mathbf{R}^{d_f\times d_k}$, $\mathbf{W}_j^V \in \mathbf{R}^{d_f\times d_v}$, $d_q = d_k$. Eq.(2) is the output for one head and there are several attention heads in order to cover different attention distributions:

$$\text{head}_j = \text{Attention}\left(\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j\right), \quad (3)$$

$$\text{MultiHead}(i) = \text{concat}\left(\text{head}_1, \ldots, \text{head}_h\right)\mathbf{W}^O \quad (4)$$

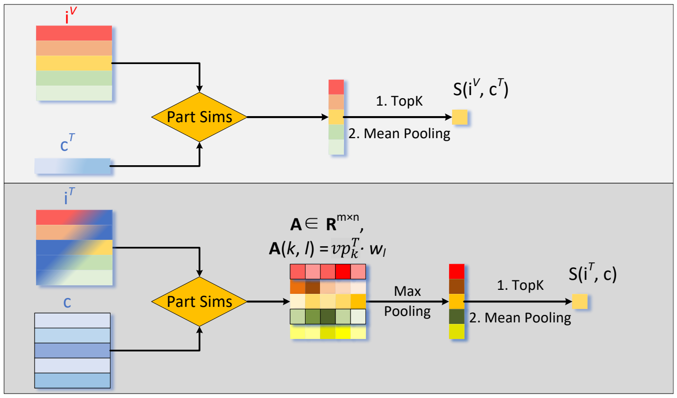


Fig. 3: Our local similarity aggregation method. Top: the method for our Intra-PAM. Bottom: the method for the I2T branch in our Inter-PAM. And the T2I branch is similar.

where $\mathbf{W}^O \in \mathbf{R}^{hd_v\times d_k}$ and h is the number of different heads. Then, the position wise feed-forward sublayer is added to further refines each vector outputted by the multi-head attention sublayer separately:

$$FFN(x_k) = \text{ReLu}\left(x_k\mathbf{W}_1 + b_1\right)\mathbf{W}_2 + b_2, \quad (5)$$

where $\mathbf{W}_1 \in \mathbf{R}^{d_k\times d_x}$, $\mathbf{W}_2 \in \mathbf{R}^{d_x\times d_k}$, $b_1 \in \mathbf{R}^{1\times d_x}$, $b_2 \in \mathbf{R}^{1\times d_k}$, and x_k is obtained via MultiHead(i) as in Eq. (4). Here, we denote $\{x_1; \cdots; x_k; \cdots; x_m\} = \text{MultiHead}(i)$, where $x_m \in \mathbf{R}^{1\times d_k}$. Besides those enumerated above, a residual connection followed by a layer normalization is applied for each sublayer. We use a transformer encoder unit as our phrase-attention module to construct image regions. The visual phrases output is denoted as:

$$i^V = \{r_1^V; r_2^V; \cdots; r_m^V\}. \quad (6)$$

Here, $r_k^V = FFN(x_k)$. The ℓ_2 normalization is applied to each r_k^V in the refined feature set i^V for the purpose of calculating the cosine similarity.

Textual Phrase Construction. Given the sentence representation $c = \{w_1; w_2; \cdots; w_n\}$, while each w_k can be seen as a word, we use a 1-d CNN [68] to further extract the local phrase information by considering the adjacent context information inherent in the sentence. To be more specific, we follow [24] to use three window sizes to capture different phrase information, named uni-gram, bi-gram and tri-gram. The output using the window size l for the k-th word is:

$$p_{l,k} = \text{ReLU}\left(\mathbf{W}_l w_{k:k+l-1} + b_l\right), \quad l = 1, 2, 3, \quad (7)$$

where $\mathbf{W}_l$ is the convolution filter matrix, and b_l is the bias. Then a max-pooling operation across all word locations is executed as follows:

$$q_l = \max\left\{p_{l,1}, \ldots, p_{l,n}\right\}, \quad l = 1, 2, 3. \quad (8)$$

In the end, we concatenate q_1, q_2, q_3 together and pass it through a fully-connected layer followed by the L_2 normalization to get a global sentence representation c^T:

$$c^T = \text{LayerNorm}\left(\mathbf{W}_e \,\text{concat}\left(q_1, q_2, q_3\right) + b_e\right), \quad (9)$$

where $\mathbf{W}_e \in \mathbf{R}^{d_k\times 3d_k}$ and $b_e \in \mathbf{R}^{d_k\times 1}$.

Similarity Measurement. Given the image phrase-aware region representation $i^V = \{r_1^V; r_2^V; \cdots; r_m^V\}$ and the sentence phrase-aware global representation c^T, we can calculate the cosine similarity between each r_k^V and c^T. As shown in the top half of Figure 3, the global similarity between i^V and c^T is the average of the top K similarities as defined in Eq. (10).

$$S(i^V, c^T) = \frac{1}{K}\sum_{a=1}^{K} \mathrm{topK_k}(r_k^V \cdot c^T). \tag{10}$$

Here, Eq.(10) calculates the similarities between the most informative textual phrase c^T and its K best-matched visual phrases, which can reduce unwanted noise as not all visual phrases correspond to this textual phrase. We demonstrate its effectiveness in subsequent ablation studies.

C. Inter-Modal Phrase-Attention Module

A great deal of work has demonstrated the importance of inter-modal interactions in image-sentence matching [1], [2], [21], [22]. Guided by these efforts and our motivation, we designed an Inter-Modal Phrase-Attention Module to achieve the inter-modal alignment with respect to visual phrases and textual phrases in a decoupled manner. The details of our Inter-PAM are illustrated in Figure 2, which shows the Inter-PAM can be divided into two branches including a Text to Image (T2I) Branch, and an Image to Text (I2T) Branch. The T2I Branch is used to model text-to-image interactions for T2I matching, while the I2T Branch models image-to-text interactions for I2T matching. Unlike existing works, we perform I2T and T2I subtasks in an asymmetric way, achieving them in the textual guided embedding space and visual guided embedding space respectively. We argue that this will help to release the trade-off between I2T matching and T2I matching and lead to better performance. We also demonstrate its effectiveness in ablation studies.

Image to Text Branch. As shown in the upper part of the inter-PAM region in Figure 2, the I2T branch includes a transformer decoder, which is the key module for the inter-modal interaction. The transformer decoder has almost the same structure as the encoder, except for an extra multi-head attention sublayer, which is suitable for our inter-modal interaction operation. Specifically speaking, given the representation of an image $i = \{r_1; r_2; \cdots; r_m\}$ extracted by Faster R-CNN, we feed it into the first multi-head attention layer, so every r_k in i can attend to each other, and can refine its own representation with multi-head attentions. We denote the output as $i^P = \{p_1; p_2; \cdots; p_m\}$, indicating each p_k as a visual phrase. i^P is then fed into the second multi-head attention sublayer to calculate queries, while c is also fed into this sublayer to calculate key-value pairs as in Eq.(1):

$$\mathbf{Q}_j^V = i^P \mathbf{W}_j^Q, \mathbf{K}_j^T = c\mathbf{W}_j^K, \mathbf{V}_j^T = c\mathbf{W}_j^V. \tag{11}$$

Then we calculate the multi-head attention as described in Eq.(2), Eq.(3) and Eq.(4) using $\mathbf{Q}_j^V$ as queries and $\mathbf{K}_j^T, \mathbf{V}_j^T$ as key-value pairs. In this process, each p_k in i^P can attend to words in c ($w_k, k = 1, 2, \cdots, n$), and find its own best-matched textual phrase to reconstruct their own representations. In other words, we use the transformer decoder to map image embeddings into textual space. Note that the mapping process also plays the role of the alignment from visual phrases to textual phrases. The output is denoted as:

$$i^T = \{vp_1^T; vp_2^T; \cdots; vp_m^T\}, \tag{12}$$

which is mapped to the textual semantic space represented by c. We then calculate the I2T similarities for the (i^T, c) pair. As shown in the bottom half of Figure 3, we first calculate a cosine similarity matrix between each normalized visual phrase vp_k^T in i^T and each normalized w_l in c. The output can be denoted as:

$$\mathbf{A} \in \mathbf{R}^{m\times n}, \mathbf{A}(k,l) = vp_k^T \cdot w_l, \tag{13}$$

and the I2T similarity is calculate by:

$$S(i^T, c) = \frac{1}{K}\sum_{a=1}^{K} \mathrm{topK}_k[\max_l \mathbf{A}(k,l)]. \tag{14}$$

That is, we first choose the most matched textual word for each visual phrase and select the top K best-matched (visual phrase, word) pairs. Then, we average their similarities as the global I2T similarity. This operation arises from similar motivation as Eq.(10), and its effectiveness is discussed latter.

Text to Image Branch. This branch is the same as the image to text branch, except for different queries and key-value pairs. We compute the similarity for T2I matching and denote it as $S(i, c^V)$ in this branch.

D. Alignment Objective and Training Strategy

According to the above discussions, we can get 3 similarity pairs: $S(i^V, c^T)$, $S(i^T, c)$, $S(i, c^V)$. Then our model can be trained with a bi-directional triplet ranking loss with hard negatives:

$$\begin{aligned}\mathcal{L} = &\max[0, M - S(I,T) + S(I,\hat{T})] \\ &+ \max[0, M - S(I,T) + S(\hat{I},T)].\end{aligned} \tag{15}$$

Here, M is the margin, (I, T) is the matched image-sentence pair, and $\hat{I}$ and $\hat{T}$ are hard negatives in a mini-batch. We train our Intra-PAM and Inter-PAM respectively. Specifically, we firstly train our Intra-PAM, and then train our Inter-PAM with our trained Intra-PAM frozen. So we first use Eq.(15) with $S(i^V, c^T)$, then with the $(S(i^T, c), S(i, c^V))$ pair, where $S(i^T, c)$ and $S(i, c^V)$ are substituted into the first and second lines of Eq.(15), respectively. We find that training our two modules respectively yielded better performance than end-to-end training. We think this is because training the two modules respectively makes them more complementary. The discussion on complementarity can refer to the visualization of the top K regions selected by the two modules in the experiment section. When inferring, we use $W_1 \cdot S(i^V, c^T) + W_2 \cdot S(i^T, c)$ for the I2T task and $W_1 \cdot S(i^V, c^T) + W_2 \cdot S(i, c^V)$ for the T2I task. The effect of each module will be discussed in ablation studies.

E. Computational Complexity

Geigle et al. [69] discuss computational complexity in cross-modal retrieval and divide cross-modal retrieval methods into “embedding-based” and “attention-based”. A typical efficient “embedding-based” method encodes images and text **separately** and then performs cross-modal retrieval, using standard

TABLE I: Comparison results of the Image-Text Retrieval on Flickr30K and COCO 1K test sets in terms of Recall@K(R@K). *: Models with external knowledge, †: ensemble results of two models. The top two results have been highlighted in bold.

Data Split	Flickr30K 1K Test							COCO 5-fold 1K Test						
Eval Task	IMG→ TEXT			TEXT→ IMG				IMG→ TEXT			TEXT→ IMG			
Method	R@1	R@5	R@10	R@1	R@5	R@10	RSUM	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
DAN [3]	55.0	81.8	89.0	39.4	69.2	79.1	413.5	-	-	-	-	-	-	-
VSE++ [43]	52.9	80.5	87.2	39.6	70.1	79.5	409.8	64.6	90.0	95.7	52.0	84.3	92.0	478.6
GXN [50]	56.8	-	89.6	41.5	-	80.1	-	68.5	-	97.9	56.6	-	94.5	-
SCO [23]	55.5	82.0	89.3	41.1	70.5	80.1	418.5	69.9	92.9	97.5	56.7	87.5	94.8	499.3
CAMP [22]	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
SCAN† [1]	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
SAEM [24]	69.1	91.0	95.1	52.4	81.1	88.1	476.8	71.2	94.1	97.7	57.8	88.6	94.9	504.3
IMRAM [42]	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
MMCA [26]	74.2	92.8	96.4	54.8	81.4	87.8	487.4	74.8	95.6	97.7	61.6	89.8	95.2	514.7
Unicoder-VL [36]	73.0	89.0	94.1	57.8	82.2	88.9	485.0	75.1	94.3	97.8	63.9	**91.6**	**96.5**	519.2
WCGL [55]	74.8	93.3	96.8	54.8	80.6	87.5	487.8	75.4	95.5	**98.6**	60.8	89.3	95.3	514.9
DCPA														
Intra-PAM [GRU]	61.8	87.2	92.6	46.2	75.2	83.7	446.7	75.6	95.2	97.4	59.5	87.5	93.6	508.8
Full-Model [GRU]	67.1	89.5	94.6	51.4	79.0	86.3	467.9	76.8	95.8	97.7	62.6	90.2	95.2	518.3
Intra-PAM	73.1	92.0	95.3	55.1	82.5	89.2	487.2	76.5	95.3	98.6	60.5	88.7	95.2	514.8
Full-Model	75.3	**93.7**	**97.5**	**59.9**	**85.0**	**91.1**	**502.5**	**78.9**	95.5	**98.6**	**64.7**	90.8	96.3	**524.8**
SMFEA* [57]	73.7	92.5	96.1	54.7	82.1	88.4	487.5	75.1	95.4	98.3	62.5	90.1	96.2	517.6
GSMN* [33]	**76.4**	**94.3**	**97.3**	57.4	82.3	89.0	496.8	78.4	**96.4**	**98.6**	63.3	90.1	95.7	522.5
SAN* [32]	**75.5**	92.6	96.2	**60.1**	**84.7**	**90.6**	**499.7**	**85.4**	**97.5**	**99.0**	**69.1**	**93.4**	**97.2**	**541.6**

TABLE II: Comparison results of the Image-Text Retrieval on COCO 5K test set in terms of Recall@K(R@K). *: Models with external knowledge, †: ensemble results of two models. The best results have been highlighted in bold.

Data Split	COCO 5-fold 1K Test						
Eval Task	IMG→ TEXT			TEXT→ IMG			
Method	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
VSE++ [43]	41.3	-	81.2	30.3	-	72.4	-
GXN [50]	42.0	-	84.7	31.7	-	74.6	-
SCO [23]	42.8	72.3	83.0	33.1	62.9	75.5	369.6
CAMP [22]	50.1	82.1	89.7	39.0	68.9	80.2	410.0
SCAN† [1]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
IMRAM [42]	53.7	**83.2**	**91.0**	39.7	69.1	79.8	416.5
MMCA [26]	**54.0**	82.5	90.7	38.7	69.7	80.8	416.4
DCPA	53.5	82.4	90.2	**40.4**	**71.0**	**82.0**	**419.5**

distance metrics to identify the most similar examples for each query in the target data collection via standard nearest-neighbor search. While an "attention-based" method utilises a cross-attention mechanism between examples from the two modalities to compute their similarity score, thus has to pass each text-image pair through the network. Assume there are N image-text pairs, encoding separately makes the computational complexity of the "embedding-based" approaches to be $O(2N)$, while the cross-attention mechanism makes the computational complexity of the "attention-based" approaches to be $O(N^2)$. While the latest approaches [70]–[73] focus on the "attention-based" methods as they usually achieve better performance, they comes at a prohibitive cost. A number of approaches have attempted to solve this problem [69], [74]–[76], which will not be described here due to space limitations. Our framework belongs to the "attention-based" methods, too.

F. Discussions

In our proposed model, we are able to not only utilize the Intra-PAM module to construct visual and textual phrases respectively with the intra-modal phrase-attention, but also use the well-designed Inter-PAM module to model the inter-modal phrase-level alignment. Here, the alignment is implemented by first modeling the phrases in one modality and then finding their phrase-level representations in the other modality's semantic space. Our Intra-PAM inherits from SAEM [24], yet we change the similarity measure to Eq. (10) which is more appropriate for our method utilizing the phrase-level alignment. Furthermore, SAEM [24] first pools local features into global features, then calculates the image-sentence similarity using global features. Here, the fine-grained matching may be not well as some noises could be introduced due to the improper alignment between visual and textual phrases.

IV. Experiment

In this section, we first introduce datasets, evaluation metrics and implementation details. Then, we show experimental results, including comparison with state-of-the-art methods, ablation studies and model analysis.

A. Datasets and Evaluation Metrics

Datasets. Two benchmark datasets are used in our experiments, including (1) **Flickr30K** [77]: A dateset consists of 31,783 images and 158,915 English texts. Each image is annotated with 5 texts. We follow the dataset splits as [21] and use 29,000 images for training, 1,000 images for validation, and the remaining 1,000 images for testing. (2) **MS-COCO** [78]: A large-scale image description dataset contains about 123,287 images and 123,287 × 5 = 616,435 sentences. We

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3197972

TABLE III: The impact of similarity measurement way on Flickr30K in the 1st stage of training.

Experiment	Batch	IMG→ TEXT			TEXT→ IMG			RSUM
Settings	Size	R@1	R@5	R@10	R@1	R@5	R@10	
global-ave	128	66.0	89.2	94.5	51.7	80.0	87.6	468.9
local-agg		68.8	91.0	95.7	52.9	80.7	88.3	477.4
global-ave	256	68.4	90.3	94.3	51.2	79.9	87.8	471.9
local-agg		72.0	90.9	94.7	54.6	81.7	88.7	482.6
global-ave	300	69.1	90.3	94.8	52.0	80.8	88.3	475.2
local-agg		73.1	92.0	95.3	55.1	82.5	89.2	487.2

TABLE IV: The impact of similarity measurement way and decoupled manner on Flickr30K in the 2nd stage of training.

Experiment	IMG→ TEXT			TEXT→ IMG			RSUM
Settings	R@1	R@5	R@10	R@1	R@5	R@10	
w/o decoupled	74.4	93.5	96.6	56.6	83.8	90.6	495.5
w/o local-agg	73.6	92.6	96.3	58.0	84.4	90.8	495.7
Full-model	75.3	93.7	97.5	59.9	85.0	91.1	502.5

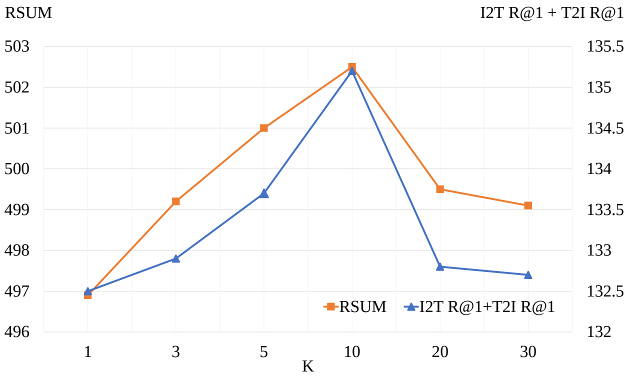


Fig. 4: Impact of K on performance. The horizontal axis is the magnitude of K, while the two vertical axes reflect the magnitude of RSUM and of the sum of the values R@1 of I2T and T2I, respectively.

follow [43], and use 113,287 images to train all models, 5,000 images for validation and another 5,000 images for testing.

Evaluation Metrics. We utlize the commonly used Recall@K (K=1,5,10) indicator as evaluation metrics to measure the performance of bi-directional retrieval tasks, where the higher Recall@K indicates better performance. Additionally, to show the overall matching performance, we also report an extra metric "RSUM", which is the sum of all the Recall values.

B. Implementation Details

We implement our method in PyTorch framework [79] with a Tesla V100-PCIE GPU. The region feature vector extracted by a bottom-up attention module [12] is 2048-dimensional, and is transformed to 256-dimensional by a fully-connect layer. As for the sentence, we use the pretrained BERT model [65] including 12 layers, 12 heads, 768 hidden units for each token. We set the pretrained BERT model fixed during training. We apply a one-layer transformer encoder with 16 heads for our Intra-PAM to construct visual phrases, and a 1d CNN with 512 filters for each filter size in Intra-PAM to construct textual phrases. In Inter-PAM, a fully-connect layer is used to transform each token in sentence to 256-dimensional, and the two transformer decoders both have one layer of 16 heads and 256 hidden units. We choose $K = 10$ for Eq.(10) and Eq.(14) and $M = 0.1$ for Eq. (15). We set $W_1 = 0.6$ and $W_2 = 0.4$ for inferring. We adopt the Adam optimizer [80] with cosine annealing with warm restart [81] as our learning rate strategy, where the initial learning rate is set to 0.0002 to train our model for 50, 30 epoches with batch-size set to 300, 84 for two stages respectively.

C. Comparison with State-of-the-art Methods

We compare our proposed model with several recently published state-of-the-art methods on the two benchmark datasets, and the results are shown in Table I and Table II.

Results on Flickr30K. As shown in Table I, our proposed DCPA achieves 502.5 for RSUM which measures the overall performance, and 75.3, 59.9 for R@1 in I2T and T2I subtasks respectively. Compared with SOTA ([26], [42], [55]) without external knowledge, DCPA is superior on all the indicators, and outperforms them by a large margin for RSUM (18.3, 15.1, 14.7). For pre-training methods like [36], we compare our results with their non-pre-training version, that is, they are trained on task-specific training data directly, without pre-training. The results show that our method exceeds [36] by 2.3 and 1.7 percentage points for R@1 in I2T and T2I subtasks respectively and also has great advantages on RSUM (17.5). When compared with methods [32], [33], [57] with external knowledge (denoted with *), DCPA is still superior to them on the overall performance indicator RSUM (502.5 vs 487.5, 496.8 and 499.7), and has considerable advantages over GSMN [33] in the T2I retrieval (59.9 vs 57.4 for R@1).

Results on MS-COCO. Table I lists the quantitative results on COCO 1K testing set by comparing with previous methods. We can find that our DCPA still surpasses other methods in most evaluation metrics, achieving 524.8 for RSUM, and 78.9, 64.7 for R@1 in I2T and T2I subtasks respectively. Besides, our method still outperforms [36] in R@1 (78.9 vs 75.1, 64.7 vs 63.9) and RSUM (524.8 vs 519.2). Compared with SOTA [26], [42], [55] without external knowledge, DCPA exceeds them by 2.2 and 3.0 percentage points for R@1 in I2T and T2I subtasks respectively. When compared with SMFEA [57] and GSMN [33] with external knowledge, DCPA outperforms it for RSUM (524.8 vs 517.6 and 522.5), and still has considerable advantages in the T2I retrieval task. The results on the MS-COCO 5K testing dataset are shown in Table II. We can see that our approach still outperforms other approaches, especially for the text to image retrieval, which is consistent with results on other different datasets.

Analysis. Compare the I2T retrieval with the T2I retrieval, our approach improves the performance of the T2I retrieval more, which is easy to explain as a sentence has a natural tree structure and is made up of multiple phrases. This makes it easier to build textual phrases than visual phrases which are constructed by several regions lacking semantic relevance. Besides, compared with GSMN, we can see that matching

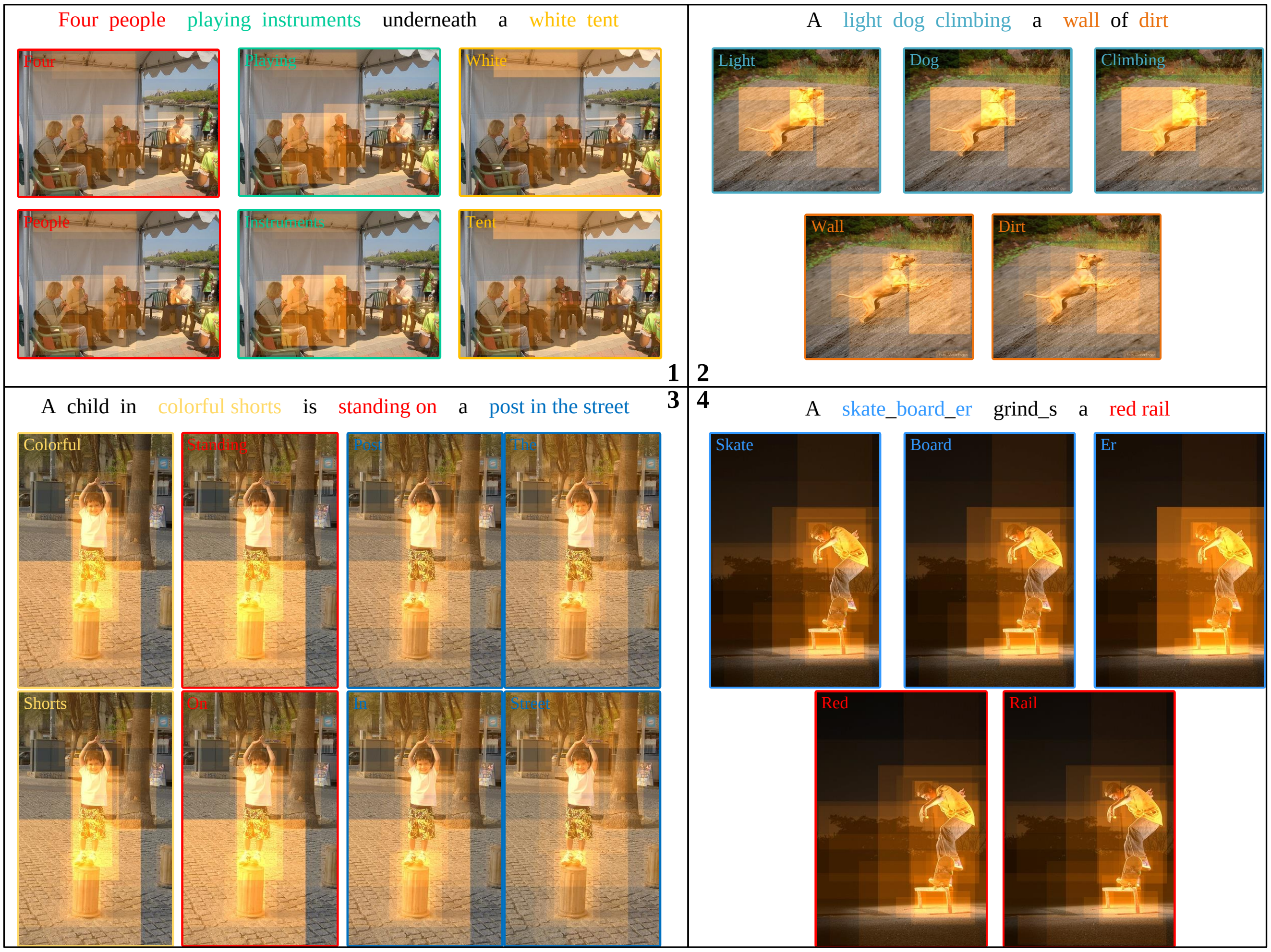

Four people playing instruments underneath a white tent
Four
Playing
White
People
Instruments
Tent
1
A light dog climbing a wall of dirt
Light
Dog
Climbing
Wall
Dirt
2
A child in colorful shorts is standing on a post in the street
3
Colorful
Standing
Post
The
Shorts
On
In
Street
4
A skate_board_er grind_s a red rail
Skate
Board
Er
Red
Rail

Fig. 5: Visualization of the phrase-attention for our Inter-PAM's T2I branch. For each sentence, we mark different textual phrases with different colors. For each textual phrase in the sentence, we show the attention map relative to the regions (bounding boxes) of the sentence's ground-truth image for each word in the phrase. We mark the corresponding word in the upper left corner of each attention map.

with phrases aggregated by our phrase-attention mechanism is no worse than matching with a well-designed graph for image sentence matching task.

D. Ablation Studies and Model Analysis

To analyze the contribution of each component and setting in our method, we perform a series of ablation studies on Flickr30K dataset and MS-COCO dataset. Our two-stage training strategy provides convenience for analyzing the effectiveness for each component. In addition, we verify the effectiveness of our similarity measurement way and their sensitivity to K (Eq.(10), Eq.(14)) . We also verify the effectiveness of our decoupled setting in the second stage of training. Moreover, in order to analyze the strengths and weaknesses of our approach, we carry out some visualization experiments of phrase-attention, top K bounding boxes selected by Eq.(10) and Eq.(14) and case study of the model.

Impact of Different Backbone. In Table I, we also show the results of changing the textual backbone from BERT [65] to a Bi-GRU. We notice that performance degradation on MS-COCO is much smaller than that on Flickr30K. We argue that this is because MS-COCO dataset is much larger than Flickr30K dataset, which makes up the performance gap between the Bi-GRU and BERT [65].

Impact of Different Modules. We first evaluate the performance of our Inter-PAM on Flickr30K and MS-COCO datasets, the results are shown in Table I. Here, Intra-PAM means that we only take advantage of our Intra-PAM to construct visual phrases and textual phrases. As shown in the table, the full model largely exceeds the network that performs matching only with Intra-PAM. This indicates the effectiveness of our full model and our elaborate Inter-PAM, and further indicates the importance of performing inter-modal interaction in this task. Besides, compared with SAEM's results [24] in Table I, our Intra-PAM achieves much better performance and the reason has been declared in Sec. III-F.

Impact of Similarity Measurements. To validate the impact of our proposed similarity measurement way (Eq.(10) and Eq.(14)) and our decoupled way to measure the I2T similarity and the T2I similarity, we conduct extensive experiments on

Flickr30K dataset. Specifically, we first verify the validity of Eq.(10) in the first stage of training, and the results are shown in Table III. In Table III, local-agg means that we use Eq.(10) to calculate the global similarity while global-ave follows SAEM to calculate the global similarity as described in Sec. III-F. We observe that the local-agg way does improve performance a lot, which is more than 10 points higher on RSUM than the global-ave way for different batch-size settings. Besides, as batch-size increases, the local-agg way's performance grows faster than the global-ave way, thus widens the performance gap between them. According to the above analysis, the local-agg way has a better performance upper bound as batch-size grows. Then we fix the best setting in the first stage of training, demonstrate the effectiveness of Eq.(14) and the decoupled similarity measure in the second stage of training, while the results are shown in Table IV. In this table, w/o decoupled means that we average the I2T similarity and the T2I similarity as one symmetric global similarity used for both I2T and T2I retrieval tasks during training, while w/o local-agg means we first average the local features as the global features and then calculate the global similarity with global features.

In general, we achieve the best results by combining these two settings which demonstrate the validity of them in the second stage of training. We also explore the impact of different sizes of K in Eq.(10) and Eq.(14) on performance, and the results are shown in Figure 4. According to Figure 4, the performance of the model shows a unimodal distribution of increasing first and then decreasing with the increase of K. This is consistent with our cognition: when K is small, the measure of similarity is not robust enough; when K is too large, a lot of background element noise is introduced. Based on the above observations, we set K to 10 as our best setting.

Visualization of phrase-attention. We visualize the phrase-attention for our Inter-PAM's T2I branch in Figure 5. It is not intuitive to visualize the textual phrase constructed by our module, so we compared the attention map of each word in the same phrase to the regions instead. As shown in Figure 5, words belonging to the same textual phrases focus on the same regions of the image, regardless of their original meanings. For example, in case 2, the query sentence consists of two main semantic parts: a climbing light dog and a dirt wall. Our Inter-PAM captures this correctly and maps them to the corresponding regions of the image respectively. In addition, we are surprised to find that our Inter-PAM can capture state of motion information by focusing on motion subjects. For instance, in case 1, the phrase "playing instruments" correctly focuses on instruments and the people who play them. We also notice some failures with phrase-attention. For example, in case 3, the phrase "colorful shorts" focuses on the wrong area in the image.

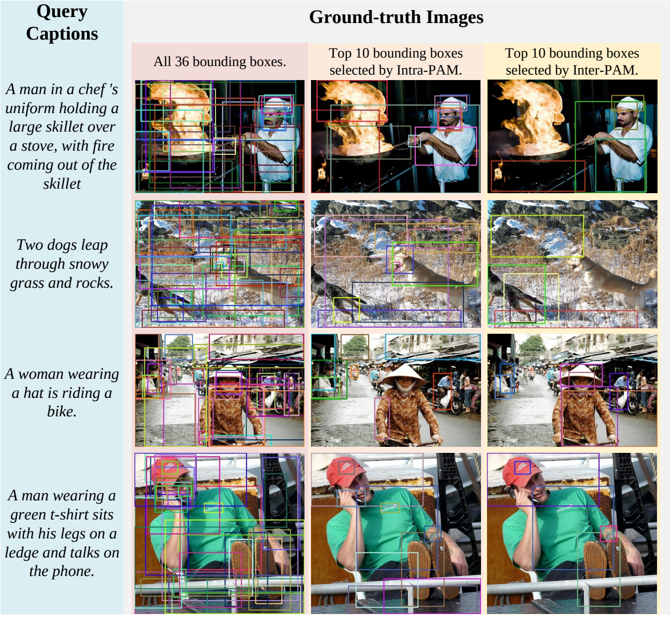
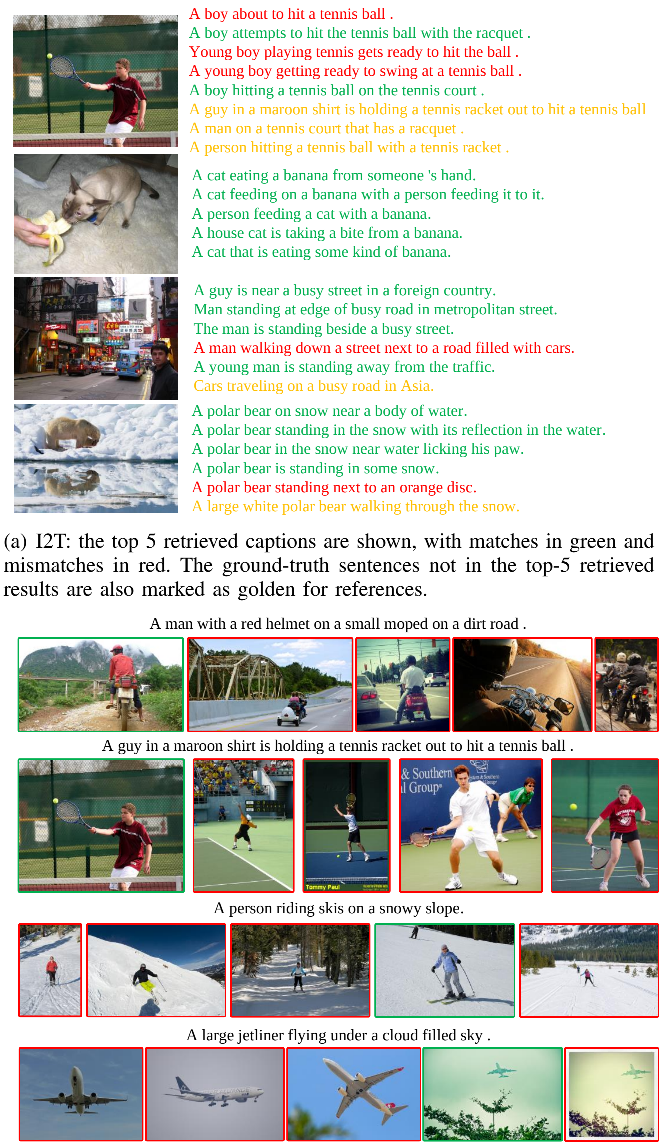

Query Captions
Ground-truth Images
All 36 bounding boxes.
Top 10 bounding boxes selected by Intra-PAM.
Top 10 bounding boxes selected by Inter-PAM.
A man in a chef 's uniform holding a large skillet over a stove, with fire coming out of the skillet
Two dogs leap through snowy grass and rocks.
A woman wearing a hat is riding a bike.
A man wearing a green t-shirt sits with his legs on a ledge and talks on the phone.

Fig. 6: The top K bounding boxes selected by our modules. The left part of this figure shows the query captions. The right part of this figure shows the ground-truth images of each caption. In the **Ground-truth Images** column of the chart, the 1st column shows all 36 bounding boxes, which is provided by pre-trained Faster R-CNN, the 2nd column shows the top 10 bounding boxes selected by our Intra-PAM using Eq. (10), and the 3rd column shows the top 10 bounding boxes selected by our Inter-PAM's T2I branch using a symmetric version of Eq. (14). Note that the 3rd column may shows fewer than 10 bounding boxes as different textual phrases may pick the same bounding box.

Visualization of the top K bounding boxes.. Figure 6 shows the top K bounding boxes selected by our modules. We observe that our two modules can complement each other by noticing different important information in the image. For example, in line 1, Inter-PAM notices the noun "stove" , which is ignored by Intra-PAM. Yet Intra-PAM notices the verb "holding", the noun phrase "a large skillet" and the noun "fire", which are ignored by Inter-PAM. While in line 2, two dogs are noticed by our two modules separately. In line 4, our Inter-PAM ignores two nouns "ledge" and "phone", which Intra-PAM succeeds in noticing. An extreme example of the complementarity of our two modules appears in the third line, where Intra-PAM fails to focus on the right area and Inter-PAM successfully notices two important concepts in the image: "woman" and "hat".

Case study. Figure 7 shows the I2T and T2I retrieval results. We can see that our approach usually retrieves the ground truth with a high rank, while some other retrieved results are also reasonable. In I2T retrieval, we can see that the wrong retrieved sentences are actually semantically similar. For example, in the 1st result, we find that 3 ground-truth sentences rank low, and they do not include the word "boy". This may be because "boy" is more specific than "guy" , "man" or "person" and can lead to better alignment. The same thing happens for the 3rd result, in which "man" is easier to be detected than "asia" and can furthermore lead to better alignment. In T2I retrieval, we find that the noise of the dataset itself leads to errors in retrieval results. For example, in the 3rd retrieval result, the top-5 results are all reasonable, but only the fourth result is the ground-truth. The 4th result in both I2T and T2I retrieval shows an interesting

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2022.3197972


A boy about to hit a tennis ball .
A boy attempts to hit the tennis ball with the racquet .
Young boy playing tennis gets ready to hit the ball .
A young boy getting ready to swing at a tennis ball .
A boy hitting a tennis ball on the tennis court .
A guy in a maroon shirt is holding a tennis racket out to hit a tennis ball
A man on a tennis court that has a racquet .
A person hitting a tennis ball with a tennis racket .
A cat eating a banana from someone 's hand.
A cat feeding on a banana with a person feeding it to it.
A person feeding a cat with a banana.
A house cat is taking a bite from a banana.
A cat that is eating some kind of banana.
A guy is near a busy street in a foreign country.
Man standing at edge of busy road in metropolitan street.
The man is standing beside a busy street.
A man walking down a street next to a road filled with cars.
A young man is standing away from the traffic.
Cars traveling on a busy road in Asia.
A polar bear on snow near a body of water.
A polar bear standing in the snow with its reflection in the water.
A polar bear in the snow near water licking his paw.
A polar bear is standing in some snow.
A polar bear standing next to an orange disc.
A large white polar bear walking through the snow.

(a) I2T: the top 5 retrieved captions are shown, with matches in green and mismatches in red. The ground-truth sentences not in the top-5 retrieved results are also marked as golden for references.

A man with a red helmet on a small moped on a dirt road .
A guy in a maroon shirt is holding a tennis racket out to hit a tennis ball .
A person riding skis on a snowy slope.
A large jetliner flying under a cloud filled sky .

(b) T2I: the top 5 retrieved images are shown, with matches in green and mismatches in red.

Fig. 7: I2T and T2I retrieval results on COCO 1K test set.

problem, that is our method fails to deal well with interfering elements. In this case, an extraneous element (“an orange disc” or “cloud”) appears in the retrieved sentence and image. However, both of them still rank higher than one of the ground-truth sentence or image. This problem is called “distraction” by Li et al. [29]. In this paper, Li et al. [29] adopts information entropy to quantify distraction and uses distraction scores to re-rank initial retrieved results. However, this approach fails to take into account the different importance of distraction in images and sentences and has made only minor improvements. Since sentences are often incomplete descriptions of images, distraction in an image just makes the results less reliable, while distraction in a sentence always means an incorrect retrieval result, whether the image or the sentence is used as a retrieval result or a query. We find this problem to be a common one in image-sentence matching and has not been carefully studied. We will consider this problem in future studies.

V. Conclusion

In this paper, we propose a cross-modal phrase-attention network for image and sentence matching, which performs matching on visual phrases and textual phrases. Here, a novel decoupled manner is designed to execute I2T matching and T2I matching in the textual guided embedding space and the visual guided embedding space respectively, which not only reduces the difficulty of optimization, but also releases the trade-off between I2T and T2I subtasks. Extensive experiments demonstrate the effectiveness and superiority of our model.

In addition, we show the inadequacies of our framwork, including high computational complexity and inability to handle distraction. In the future, we will explicitly explore the solutions to distraction problem in fine-grained cross-modal matching. Moreover, for higher computational efficiency, it shall be necessary for us to explore a method which can model phrase-level matching without inter-modal interactions.

VI. Acknowledgements

This work was partially supported by the National Nature Science Foundation of China (62022078, 12150007, 62021001), National Defense Basic Scientific Research Program (JCKY2020903B002), and University Synergy Innovation Program of Anhui Province No. GXXT-2019-025.

References

[1] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *ECCV*, 2018, pp. 201–216.

[2] Y. Huang, W. Wang, and L. Wang, “Instance-aware image and sentence matching with selective multimodal lstm,” in *CVPR*, 2017, pp. 2310–2318.

[3] H. Nam, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *CVPR*, 2017, pp. 299–307.

[4] L. Wang, Y. Li, J. Huang, and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 394–407, 2018.

[5] T. Dutta and S. Biswas, “Generalized zero-shot cross-modal retrieval,” *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5953–5962, 2019.

[6] Y. Zhang, W. Zhou, M. Wang, Q. Tian, and H. Li, “Deep relation embedding for cross-modal retrieval,” *IEEE Transactions on Image Processing*, vol. 30, pp. 617–627, 2020.

[7] M. Meng, H. Wang, J. Yu, H. Chen, and J. Wu, “Asymmetric supervised consistent and specific hashing for cross-modal retrieval,” *IEEE Transactions on Image Processing*, vol. 30, pp. 986–1000, 2020.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, 2015, pp. 2048–2057.

[9] W. Zhao, X. Wu, and J. Luo, “Cross-domain image captioning via cross-modal retrieval and model adaptation,” *IEEE Transactions on Image Processing*, vol. 30, pp. 1180–1192, 2021.

[10] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, and D. Tao, “Rethinking diversified and discriminative proposal generation for visual grounding,” *arXiv preprint arXiv:1805.03508*, 2018.

[11] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” *arXiv preprint arXiv:1606.01847*, 2016.

[12] 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 *CVPR*, 2018, pp. 6077–6086.

[13] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6281–6290.

[14] P. Gao, Z. Jiang, H. You, P. Lu, S. C. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra-and inter-modality attention flow for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6639–6648.

[15] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[16] S. Kaya and E. Vural, "Learning multi-modal nonlinear embeddings: Performance bounds and an algorithm," *IEEE Transactions on Image Processing*, vol. 30, pp. 4384–4394, 2021.

[17] Y. Huang and L. Wang, "Acmm: Aligned cross-modal memory for few-shot image and sentence matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5774–5783.

[18] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan, "Position focused attention network for image-text matching," *arXiv preprint arXiv:1907.09748*, 2019.

[19] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1979–1988.

[20] Y. Huang, Y. Long, and L. Wang, "Few-shot image and sentence matching via gated visual-semantic embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8489–8496.

[21] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[22] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5764–5773.

[23] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *CVPR*, 2018, pp. 6163–6171.

[24] Y. Wu, S. Wang, G. Song, and Q. Huang, "Learning fragment self-attention embeddings for image-text matching," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, pp. 2088–2096.

[25] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4654–4662.

[26] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 941–10 950.

[27] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," in *Advances in Neural Information Processing Systems*, 2019, pp. 6850–6860.

[28] P. Gao, H. You, Z. Zhang, X. Wang, and H. Li, "Multi-modality latent interaction network for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5825–5835.

[29] Y. Li, D. Zhang, and Y. Mu, "Visual-semantic matching by exploring high-order attention and distraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 786–12 795.

[30] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3536–3545.

[31] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[32] Z. Ji, H. Wang, J. Han, and Y. Pang, "Saliency-guided attention network for image-sentence matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5754–5763.

[33] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 921–10 930.

[34] 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, pp. 13–23.

[35] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[36] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training." in *AAAI*, 2020, pp. 11 336–11 344.

[37] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Learning universal image-text representations," *arXiv preprint arXiv:1909.11740*, 2019.

[38] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," *arXiv preprint arXiv:1908.05054*, 2019.

[39] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7464–7473.

[40] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[41] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020.

[42] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 655–12 663.

[43] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[44] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[45] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *CVPR*, 2015, pp. 4437–4446.

[46] Y. Peng and J. Qi, "Cm-gans: cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1, p. 22, 2019.

[47] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *CVPR*, 2015, pp. 3441–3450.

[48] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

[49] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[50] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *CVPR*, 2018, pp. 7181–7189.

[51] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[52] S. Chun, S. J. Oh, R. S. De Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8415–8424.

[53] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 789–15 798.

[54] S. Yan, L. Yu, and Y. Xie, "Discrete-continuous action space policy gradient-based attention for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8096–8105.

[55] Y. Wang, T. Zhang, X. Zhang, Z. Cui, Y. Huang, P. Shen, S. Li, and J. Yang, "Wasserstein coupled graph learning for cross-modal retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1813–1822.

[56] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," *arXiv preprint arXiv:2101.01368*, 2021.

[57] X. Ge, F. Chen, J. M. Jose, Z. Ji, Z. Wu, and X. Liu, "Structured multimodal feature embedding and alignment for image-sentence retrieval," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5185–5193.

[58] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[59] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[60] K. Wen, J. Xia, Y. Huang, L. Li, J. Xu, and J. Shao, "Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2208–2217.

[61] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[62] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6995–7004.

[63] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 976–12 985.

[64] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[66] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[68] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[69] G. Geigle, J. Pfeiffer, N. Reimers, I. Vulić, and I. Gurevych, "Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval," *arXiv preprint arXiv:2103.11920*, 2021.

[70] P. Zeng, L. Gao, X. Lyu, S. Jing, and J. Song, "Conceptual and syntactical cross-modal alignment with cross-level consistency for image-text matching," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2205–2213.

[71] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-vil: Knowledge enhanced vision-language representations through scene graph," *arXiv preprint arXiv:2006.16934*, 2020.

[72] Y. Li, Y. Pan, T. Yao, J. Chen, and T. Mei, "Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network," *arXiv preprint arXiv:2101.11562*, 2021.

[73] M. Zhou, L. Zhou, S. Wang, Y. Cheng, L. Li, Z. Yu, and J. Liu, "Uc2: Universal cross-lingual cross-modal vision-and-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4155–4165.

[74] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Thinking fast and slow: Efficient text-to-visual retrieval with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9826–9836.

[75] J. Li, L. Liu, L. Niu, and L. Zhang, "Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval," *IEEE Transactions on Image Processing*, vol. 30, pp. 9193–9207, 2021.

[76] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu, "Compressing visual-linguistic model via knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1428–1438.

[77] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[78] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[79] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[81] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

Zhangxiang Shi received the bachelor's degree in Electronic Engineering and Information Science from University of Science and Technology of China, Hefei, China, in 2020. Currently, he is a Master degree candidate in Information and Communication Engineering, University of Science and Technology of China. His current research interests include computer vision, natural language processing and machine learning, especially image-text retrieval, video-text retrieval and video boundary detection.

Tianzhu Zhang (M'11) received the bachelor's degree in communications and information technology from Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. Currently, he is a Professor at the Department of Automation, School of Information Science, University of Science and Technology of China. His current research interests include computer vision and multimedia, especially action recognition, object classification, object tracking, and social event analysis.

Xi Wei received the bachelor's degree in Department of Automachine, University of Science and Technology of China, Hefei, Anhui, China, in 2017. He is currently pursuing the Ph.D. degree at the University of Science and Technology of China. His research interests include multimedia analysis and computer vision, especially deep learning, cross modal tasks.

Feng Wu (M'99-SM'06-F'13) received the B.S. degree in Electrical Engineering from XIDIAN University in 1992. He received the M.S. and Ph.D. degrees in Computer Science from Harbin Institute of Technology in 1996 and 1999, respectively. Now he is a professor in University of Science and Technology of China and the dean of School of Information Science and Technology. Before that, he was principle researcher and research manager with Microsoft Research Asia. His research interests include image and video compression, media communication, and media analysis and synthesis. He has authored or co-authored over 200 high quality papers (including several dozens of IEEE Transaction papers) and top conference papers on MOBICOM, SIGIR, CVPR and ACM MM. He has 77 granted US patents. His 15 techniques have been adopted into international video coding standards. As a co-author, he got the best paper award in IEEE T-CSVT 2009, PCM 2008 and SPIE VCIP 2007. Wu has been a Fellow of IEEE. He serves as an associate editor in IEEE Transactions on Circuits and System for Video Technology, IEEE Transactions on Multimedia and several other International journals. He got IEEE Circuits and Systems Society 2012 Best Associate Editor Award. He also serves as TPC chair in MMSP 2011, VCIP 2010 and PCM 2009, and Special sessions chair in ICME 2010 and ISCAS 2013.

Yongdong Zhang (M'09-SM'13) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002. He is currently a Professor at the University Of Science And Technology Of China. He has authored more than 100 refereed journal and conference papers. His current research interests include multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology. Prof. Zhang serves as an Editorial Board Member of Multimedia Systems Journal and Neurocomputing. He was the recipient of the Best Paper Award in PCM2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate in ICME 2011.