Cross-modal Semantic Enhanced Interaction for Image-Sentence Retrieval

Xuri Ge[1], Fuhai Chen[2]*, Songpei Xu[1], Fuxiang Tao[1], Joemon M. Jose[1]
[1]School of Computing Science, University of Glasgow, Glasgow, UK.
[2]Department of Computer Science, The University of Hong Kong, Hong Kong, China.
x.ge.2@research.gla.ac.uk, chenfuhai3c@163.com, s.xu.1@research.gla.ac.uk,
f.tao.1@research.gla.ac.uk, Joemon.Jose@glasgow.ac.uk

Abstract

*Image-sentence retrieval has attracted extensive research attention in multimedia and computer vision due to its promising application. The key issue lies in jointly learning the visual and textual representation to accurately estimate their similarity. To this end, the mainstream schema adopts an object-word based attention to calculate their relevance scores and refine their interactive representations with the attention features, which, however, neglects the context of the object representation on the inter-object relationship that matches the predicates in sentences. In this paper, we propose a Cross-modal Semantic Enhanced Interaction method, termed **CMSEI** for image-sentence retrieval, which correlates the intra- and inter-modal semantics between objects and words. In particular, we first design the intra-modal spatial and semantic graphs based reasoning to enhance the semantic representations of objects guided by the explicit relationships of the objects' spatial positions and their scene graph. Then the visual and textual semantic representations are refined jointly via the inter-modal interactive attention and the cross-modal alignment. To correlate the context of objects with the textual context, we further refine the visual semantic representation via the cross-level object-sentence and word-image based interactive attention. Experimental results on seven standard evaluation metrics show that the proposed CMSEI outperforms the state-of-the-art and the alternative approaches on MS-COCO and Flickr30K benchmarks.*

1. Introduction

Image-sentence retrieval aims at retrieving the most relevant images (or sentences) given a query sentence (or image), which involves the cross-over study on computer vision and neural language processing [11, 37, 9, 20, 2]. Due to its broad applications, such as multimedia analysis, multimedia search, album management, and medical image retrieval, image-sentence retrieval has aroused the widespread research attention. The key issue of image-sentence retrieval lies in jointly learning the visual and textual representations to guarantee their similarity between the matched image and sentence.

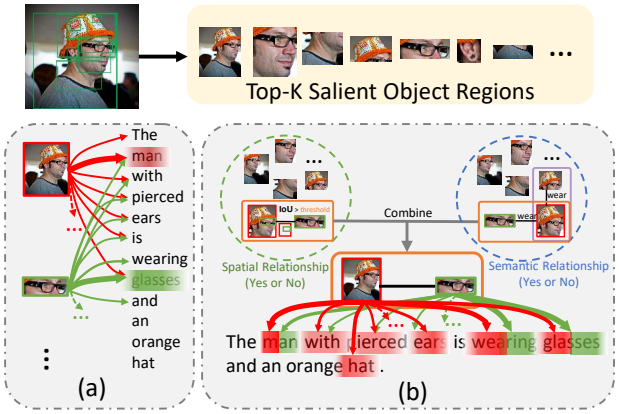

Figure 1. Illustration of two cross-modal semantic interaction schemas (only show image-to-sentence retrieval for clarity). (a) traditional schema: the correlated objects (*e.g. man* and *glass*) are hardly attended to their common word (*e.g. wear*) with high relevance score (thick arrow), (b) our CMSEI method: the relationships between the correlated objects are integrated into the region features of these objects, whose common word is with high relevance score. Note that the semantic relationships are detected via scene graph and judged by whether there is a predicted label (*e.g. wear*) with high confidence.

To this end, existing works mainly adopt two schemas to learn the visual and textual representations, *i.e.* modality-independent representation learning [11, 17, 37, 9, 49, 40, 12, 4, 5] and the cross-modal semantic interaction [20, 13, 24, 31, 47, 26] Specially, on one hand, the modal-independent representation learning has been widely studied due to its high retrieval efficiency. For instance, [11, 37, 9, 49] optimized a joint embedding space by minimizing the distance of the visual and textual global features, which are directly extracted from the whole image and the

*Corresponding author

full sentence via Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) respectively. Several recent works [13, 4, 5] extracted the visual and textual local features from object regions and words and integrated them as a whole respectively before projecting these two features into a common latent space. However, due to the lack of the deep semantic correlation on the fine-grained fragments, these methods are limited on the retrieval accuracy.

On the other hand, cross-modal semantic interaction is proposed to boost the retrieval performance by learning the accurate visual-textual semantic relevance between the fragments of image and sentence [20, 13, 24, 31, 47], as shown in Figure 1 (a). For instance, SCAN [20] attended object regions to each word to generate the text-aware visual features for sentence-to-image matching and conversely for image-to-sentence matching. Although achieving the significant improvement, these methods neglects the fact that little inter-object relationships are reflected in the object representations compared to the strong context of the textual structure, which leads to a feeble role of visual semantic during image-sentence matching.

To deal with the above problem, it's intuitive to put forward two straightforward solutions to cooperate visual semantic representation with the inter-object relationship. On one hand, the object region features can be concatenated with the feature of the inter-object relationship detected by an off-the-shelf detector [43, 41, 35, 1]. However, such method has three defects that affect the retrieval performance: (i) it's hard to keep a structured correlation among the objects and their relationships in a multi-layer network without continuous correlation guidance; (ii) the detected relationship labels, trained on different dataset, bring about the extra recognition error; and (iii) it's not an end-to-end framework. On the other hand, the inter-object relationships can be utilized for object representation enhancement via the graph-based modeling. The representative solutions are in two folds. First, following [38, 29], the relationships are detected guided by the scene graph and their label-based features are aggregated with the object region features to feed the Graph Convolution Networks (GCNs). However, such methods still suffer from the aforementioned (ii) and (iii) as revealed in [26]. Second, following [8, 5], the relationships are implicitly reflected via the fully-connected GCNs where the object region features are input as graph nodes, nevertheless, leaving the relationship information weak and ambiguous that effects the object discrimination. Therefore, it's natural for us to consider an integrated structured modeling that captures the explicit information of the inter-object relationships to enhance the object representation. As manifested in Figure 1 (b), by explicitly constructing the inter-object relationship, it's easier compared to 1 (a) for the correlated object (*e.g.* regions of *head* and *glasses*) to obtain the high relevance with the correlated predicate words (*e.g.* word *wear*).

Driven by the above consideration, we propose a novel cross-modal semantic enhanced interaction method for image-sentence retrieval, termed *CMSEI*, which correlates the intra- and inter-modal semantics between objects and words. For the intra-modal semantic correlation, the inter-object relationships are explicitly reflected on the spatially relative positions and the scene graph guided potential semantic relationships among the object regions. We then propose a relationship-aware GCNs model (termed *R-GCNs*) to enhance the object region representations with their relationships, where the graph nodes are object region features and the graph structures are determined by the inter-object relationships, *i.e.* each edge connection in the graph adjacency matrices rely on whether there is a relationship with high confidence. Different from [38, 29], intra-modal semantic correlation in CMSEI minimizes the error interference from the detection and maximizes the feasibility of the end-to-end representation learning. For the inter-modal semantic correlation, the semantic enhanced representations of words that undergo a fully-connected GCNs model, as well as the semantic enhanced representations of object regions are attended alternatively in the inter-modal interactive attention, where the object region features are attended to each word to refine its feature and conversely the word feature are attended to each object region to refine its feature. To correlate the context of objects with textual context, we further refine the representations of object regions and words via cross-level object-sentence and word-image based interactive attention. The intra-modal semantic correlation, inter-modal semantic correlation, and the similarity-based cross-modal alignment are jointly executed to further enhance the cross-modal semantic interaction.

The contributions of this paper are as follows: (1) We explore an intra-modal semantic enhanced correlation to explicitly utilize the inter-object spatially relative positions and inter-object semantic relationships guided by scene graph, and propose a relationship-aware GCNs model (R-GCNs) to enhance the object region features with their relationships. This module mitigates the error interference from the detection and enables the end-to-end representation learning. (2) We propose a cross-modal semantic enhanced interaction method (CMSEI) to unite the intra-modal semantic correlation, inter-modal semantic correlation, and the similarity-based cross-modal alignment to simultaneously model the semantic correlations on three grain levels, *i.e.* intra-fragment, inter-fragment, inter-instance. Especially, cross-level interactive attention is proposed to model the correlations between the fragments and the instance. (3) The proposed CMSEI is sufficiently evaluated with extensive experiments on MS-COCO and Flickr30K benchmarks. The results in seven standard evaluation metrics

demonstrate the superiority of the proposed CMSEI, where CMSEI achieves the state-of-the-art on the most of metrics.

2. Related Work

The key issue of the image-sentence retrieval is measuring the visual-textual similarity between an image and a sentence. It can be divided into two main kinds: modality-independent representation retrieval and cross-modal interaction retrieval. CMSEI belongs to the latter one.

Modality-independent representation retrieval. Most earlier works [11, 19, 27, 10, 34, 37, 36] used independent processing of images and sentences within two branches to obtain a holistic representation of images and sentences. Some works [11, 19, 37, 50] directly extracted the features of two modalities from the whole image via CNNs and from the full sentence via RNNs. Inspired by the detection of object regions, many studies [17, 16] started to use the pre-extracted salient object region features to represent images. And fine-grained region-level image features and word-level text features are constructed and aligned within the modalities, respectively. For instance, DVSA in [16] first adopted R-CNN to detect salient objects and inferred latent alignments between word-level textual features in sentences and region-level visual features in images. Furthermore, to take full advantages of high-level objects and words semantic information, many recent methods [28, 42, 12, 8, 22] exploited the relationships between the objects and words to help the global embedding of images and sentences, respectively. For instance, [21, 22] proposed to incorporate the semantic relationship information into visual and textual features by performing object or word relationship reasoning by GCNs.

Cross-modal interaction retrieval. Another popular retrieval schemes exploit the fine-grained cross-modal interactions [28, 20, 15, 30, 48, 3, 31] to improve the visual-textual semantic alignments. For instance, [31] proposed a method for modeling complex dynamic modal interactions based on a three-layer structure with four basic cells per layer. Recently, some works [21, 44, 25, 38, 26] employed GCNs to improve the interaction and integrate different item representations by a learned graph. Liu *et al.* [25] proposed to learn correspondence of objects and relations between modalities by two different visual and textual reasoning graphs, which is difficult to unify the two modal structures for precise pairing. Long *et al.* [26] also proposed two-modal graphs to help the interactions between modalities, however, the post-interaction concatenation did not substantially improve interactions and additionally introduced word label noise from the scene graph. And some works [38, 29, 26] also encoded the word labels from the detected visual scene graphs causing ambiguity, due to the effect of cross-domain training.

3. Approach

Figure 2 shows the overall pipeline of our proposed CMSEI for image-sentence retrieval. In this section, we will describe the detailed structure of CMSEI.

3.1. Multi-modal Feature Representations

Visual representations. To better represent the salient objects and attributes in images, we take advantage of bottom-up-attention model [1] to extract top-K saliently sub-region features $\hat{I} = \{I_j\}$, $I_j \in \mathbb{R}^{2048}$, based on the category confidence score in an image. Afterward, a fully connected (FC) layer with the parameter $W^o \in \mathbb{R}^{2048\times D_v}$ is used to project these feature vectors into a D_v-dimensional space. Finally, these projected object region features $V = \{v_1, \cdots, v_K\}$, $v_j \in \mathbb{R}^{D_v}$, are taken as initial visual representations without semantic enhancement.

Textual representations. For sentence texts, we follow the recent trends in the community of Natural Language Processing and utilize pre-trained BERT [7] model to extract word-level textual representations. Similar to visual features processing, we also utilize FC layers to project the extracted word features into a D_t-dimensional space, denoted as $T = [t_1, t_2, \cdots, t_N]$, $t_j \in \mathbb{R}^{D_t}$, with length N.

To facilitate cross-modal interaction and embedding space consistency, the projected dimensions are same (D_v=D_t) for visual and textual representations. For subsequent local-global (image-word/sentence-object) inter-modal interaction and final cross-modal similarity calculation, we use average-pooling operation to obtain the global image feature $\bar{V}$ for sentence-to-image and the global sentence feature $\bar{T}$ for image-to-sentence.

3.2. Intra-modal Relationship Enhancement

Explicit visual spatial graph. Since features from the top-K candidate object regions are used for representing the image information, this leads to some regions with semantic overlap but with minor positional bias. In addition, study [5] indicated that the regions with larger Intersection over Union (IoU) as potentially more closely. To this end, following [5], we also construct explicit spatial non-fully connected graph $G^s = (V, E^s)$ for each image. The semantic similarities and spatial IoUs between sub-regions are combined to represent the adjacency matrix $A^s \in \mathbb{R}^{K\times K}$ as edges for spatial graphs. In particular, if the IoU_{ij} of the i-th region and the j-th region exceeds the threshold μ, the semantic similarity between them is treated as a weighted edge A^s_{ij}, otherwise it is 0. The pairwise semantic similarity is updated and calculated between regions as: $Sims = (W^v_\varphi V)^T (W^v_\phi V)$ (W^v_φ and W^V_ϕ denote the mapping parameters). For simplicity, we do not explicitly represent the bias term in our paper.

Explicit visual semantic relationship graph. Different from existing approaches [21, 5] based on implicit relation-

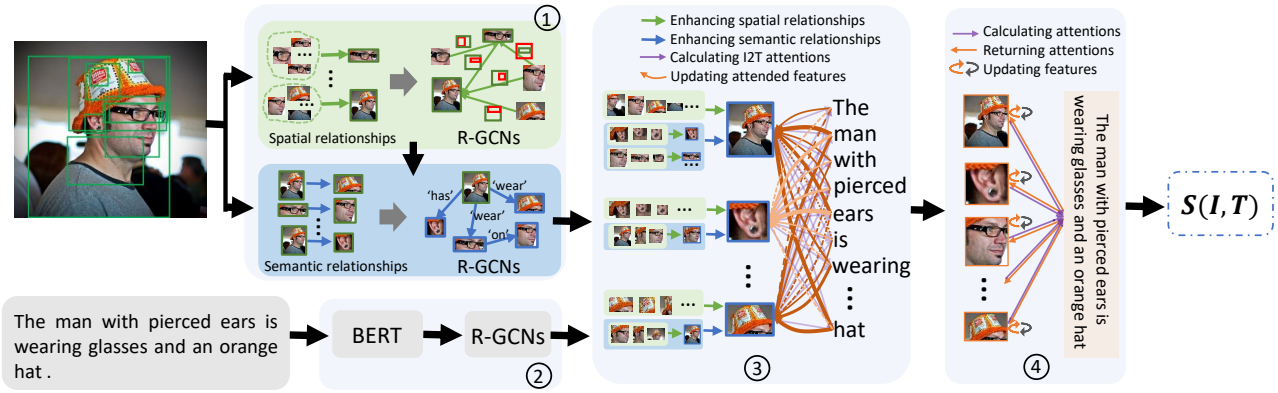

Figure 2. The overall framework (image-to-sentence version) of CMSEI. In intra-modal semantic correlation (①), two relationship-aware GCNs are constructed to respectively integrate the explicit spatial and semantic relationships between each two objects into their region representations by changing the relationship-determined graph adjacency matrices. In ②, a pre-trained BERT model is used to obtain high-level semantic embedding features of words, which are then fed to GCNs to enhance them with the context. In inter-modal semantic correlation (③ and ④), the visual and textual semantic features are further enhanced via object-word interactive attention and the visual semantic representation is refined via the cross-level object-sentence and word-image based interactive attention. Visual and textual semantic similarity is finally estimated for the cross-modal alignment.

ship graph reasoning, scene graphs have well-defined object relationships, which can overcome the disadvantage of fusing redundant information. Unlike approaches [38, 26] based on scene-graph enhancement, we do not encode the word labels predicted by the pre-trained visual scene-graph generator, like [46]. We consider word labels from visual scene graphs of external models have errors and semantically different from the words in the corresponding sentences. This tends to introduce noise that corrupts the cross-modal semantic alignment. In this paper, we construct a non-fully connected semantic relationship graph $G^v = (V^s, E^v)$ between the spatially enhanced objects of each image based on the explicit relationships of the visual scene graphs. In each relationship graph G^v, the nodes indicate the object features V^s updated from spatial graph G^s and the edges indicate the existence of semantic associations, as in Figure 1 (b). Here, we construct an adjacency matrix $A^v \in \mathbb{R}^{K\times K}$ to represent these edges for each image, where A^v_{ij}=1 means i-th object is associated with j-th object in the semantic relations extracted by a pre-trained visual scene-graph generator and 0 otherwise. Unlike the spatial graph, since the regions in the scene graph are already strongly correlated, we no longer exploit their semantic similarity.

Visual Feature Embedding. The currently popular Graph Convolutional Networks (GCNS) [21] with residuals are used to obtain the final object region features V^f, enhanced by updating and embedding of spatial and semantic relationship graphs, named relationship-aware GCNs (R-GCNs), as shown in Figure 2 ①. Formally,

$$V^s = (A^s V W^s_g)W_{r_1} + V, \tag{1}$$

$$V^f = ((A^v V W^v_g)W_{r_2} + V^s)W_{r_3} + V, \tag{2}$$

where $W^*_g \in \mathbb{R}^{D_v \times D_v}$ are the weight matrix of the GCN layer, W_{r_*} are the residual weights.

Implicit textual graph building and embedding. In contrast to the approaches [38, 29, 5, 26] of explicitly modeling inter-word dependencies, we construct a fully connected graph for each sentence, where the semantic features T of the words serve as nodes and the semantic similarities A^t between words serve as edges. We argue that explicit modeling of sentences tends to focus only on the words of object and relation and loses the benefit of many attribute descriptions. Similar to the visual enhancement process, as shown in Figure 2 ②, we apply GCNs [21, 22] with residuals to reason and get the final textual representations T^f with the relationship enhanced, as follows:

$$A^t = (W^t_\varphi T)^T (W^t_\phi T), \tag{3}$$

$$T^f = (A^t T W^t_g)W_{r_t} + T, \tag{4}$$

where W^t_φ and W^t_ϕ denote the mapping parameters, W_{r_t} is the residual weights, W^t_g is the weight matrix of the GCN layer.

3.3. Inter-modal Interactions

After image objects and sentence words are reinforced with semantic relationships within a modality, we apply two mainstream inter-modal interaction mechanisms to further enhance the feature representation of the target modality with attention-ware information from another modality.

Local-local inter-modal interaction. Similar to literature [20, 31], we mine attentions between image objects and sentence words to narrow the semantic gap between two modalities. As shown in Figure 2 ③, taking the image-to-sentence example (Due to space limitations and a clearer presentation), we first calculate the cosine similarities for all object-word pairs and calculate the attention weights by

a per-dimension λ-smoothed Softmax function [6], as follows:

$$c_{ij} = \frac{(v_i^f)^T t_j^f}{||v_i^f||\,||t_j^f||}, i \in [1, K], j \in [1, N], \quad (5)$$

$$\alpha_{ij} = \frac{exp(\lambda c_{ij})}{\sum_{j=1}^{N} exp(\lambda c_{ij})}, \quad (6)$$

Finally, we obtain the attended object representation $v_i^t \in V^t$ via a conditional fusion strategy [31] from correspondence attention-aware textual vector q_i^t (q_i^t=$\sum_{j=1}^{N} \alpha_{ij} t^f$), as follows,

$$v_i^t = \mathrm{ReLU}(W_1^t(v_i^f \odot \mathrm{Tanh}(W_2^t q_i^t) + W_3^t q_i^t)) + v_i^f, \quad (7)$$

where W_*^t are the mapping parameters, ReLU and Tanh are activation functions. To fully explore fine-grained cross-modal interactions, we perform the above process twice.

Local-global inter-modal interaction. As shown in Figure 2 ④, we further discover the salience of the fragments in one modality guided by the global contextual information of the other modality, which makes each fragment contains more contextual features. Specifically, for image-to-sentence, we first calculate the semantic similarity between the objects of image $V^t = \{v_1^t, \cdots, v_K^t\}$ and global textual feature $\bar{T}$. Then, we can obtain the relative importance of each object via a sigmoid function. Finally, we add residual connections between the attention-aware object features and the enhanced object features V^t, as well as the original features V. The above process can be formulated as:

$$r_i = \sigma(W^r v_i^t \odot \bar{T}), \quad (8)$$

$$v_i^o = r_i v_i^t + v_i^t + \mathrm{ReLU}(v_i), \quad (9)$$

where W^r denotes the mapping parameter. Similarly, for sentence-to-image, we enhance the word features via calculating the relative importance of each word between the words of the sentence and the global image feature $\bar{V}$.

To obtain the final match score between image and sentence, we average and normalise the final object features of the image and calculate the cosine similarity with the global sentence features.

3.4. Objective Function

In the above training process, all the parameters can be simultaneously optimized by minimizing a bidirectional triplet ranking loss [9], when aligning the image and sentence as follows:

$$\begin{aligned}\mathcal{L}_{rank}(I,T) = &\sum_{(I,\hat{T})} [\nabla - \cos(I,T) + \cos(I,\hat{T})]_+ \\ &+ \sum_{(\hat{I},T)} [\nabla - \cos(I,T) + \cos(\hat{I},T)]_+\end{aligned} \quad (10)$$

where ∇ serves as a margin constraint, $\cos(\cdot,\cdot)$ indicates cosine similarity function, and $[\cdot]_+ = \max(0, \cdot)$. Note that, (I, S) denotes the given matched image-sentence pair and its corresponding negative samples are denoted as $\hat{I}$ and $\hat{S}$, respectively.

4. Experiments

In this section, we report the results of our experiments to evaluate the proposed approach, CMSEI. We will introduce the dataset and experimental settings first. Then, CMSEI is compared with state-of-the-art image-sentence retrieval approaches quantitatively. Finally, we qualitatively analyze the results in detail.

4.1. Dataset and Evaluation Metrics

Dataset. We evaluate our proposed approach on the MS-COCO [23] and Flickr30k [45] datasets, which are the most popular benchmark datasets for image-sentence retrieval task. There are over 123,000 images in MS-COCO. Following the splits of most existing methods [39, 26, 22, 4, 31], there are 113,287 images for training, 5,000 images for validation and 5000 images for testing. On MS-COCO, we report results on both 5-folder 1K and full 5K test sets, which are the average results of 5 folds of 1K test images and the results of full 5K test set, respectively. Flickr30K contains over 31,000 images with 29, 000 images for the training, 1,000 images for the testing, and 1,014 images for the validation. Each image in these two benchmarks is given five corresponding sentences by different AMT workers.

Evaluation metrics. Following the standard evaluation protocol, we employ the widely-used recall metric, R@K (K=1,5,10) evaluation metric, which denotes the percentage of ground-truth being matched at top K results, respectively. Moreover, we report the "*rSum*" criterion that sums up all six recall rates of R@K, which provides a more comprehensive evaluation to testify the overall performance.

4.2. Implementation Details

Our model is trained on a single TITAN RTX GPU with 24 GB memory. The whole network except the Faster-RCNN model [32] is trained from scratch with the default initializer of PyTorch using ADAM optimizer [18] a mini-batch size 64. The learning rate is set to 0.0002 initially with a decay rate of 0.1 every 15 epochs. Maximum epoch number is set to 30. The margin of triplet ranking loss ∇ is set to 0.2. The threshold μ is set to 0.4. For the visual object features, Top-K (K=36) object regions are selected with the highest class detection confidence scores. The visual scene graphs are generated by Neural Motifs [46], and we use the maximum IoU to find the corresponding regions in the original Top-K salient regions. The initial dimensions of visual and textual embedding space are set to 2048 and 768 respectively, which are transformed to the same 1024-dimensional (*i.e.*, D_v= D_s=1024). The most dimensions of mapping parameters are set to 1024-dimensional.

Table 1. Comparisons of experimental results on MS-COCO 5-folds 1K test set and full 5K test set.

Method	Sentence Retrieval			Image Retrieval			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
5-folds 1K							
SCAN$^{*}_{\mathrm{ECCV'18}}$ [20]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
VSRN$^{*}_{\mathrm{ICCV'19}}$ [21]	76.2	94.8	98.2	62.8	89.7	95.1	516.8
IMRAM$^{*}_{\mathrm{CVPR'20}}$ [3]	76.7	95.6	98.5	61.7	89.1	95.0	516.6
CAAN$_{\mathrm{CVPR'20}}$ [48]	75.5	95.4	98.5	61.3	89.7	95.2	515.6
GSMN$^{*}_{\mathrm{CVPR'20}}$ [25]	78.4	96.4	98.6	63.3	90.1	95.7	522.5
CAMERA$^{*}_{\mathrm{ACMMM'20}}$ [30]	78.0	95.1	97.9	60.3	85.9	91.7	508.9
SGRAF$^{*}_{\mathrm{AAAI'21}}$ [8]	79.6	96.2	98.5	63.2	90.7	96.1	524.3
VSE$\infty_{\mathrm{CVPR'21}}$ [4]	79.7	96.4	**98.9**	64.8	91.4	96.3	<u>527.5</u>
DIME$^{*}_{\mathrm{SIGIR'21}}$ [31]	78.8	96.3	98.7	64.8	91.5	96.5	526.6
CGMN$^{*}_{\mathrm{TOMM'22}}$ [5]	76.8	95.4	98.3	63.8	90.7	95.7	520.7
VSRN++$^{*}_{\mathrm{TPAMI'22}}$ [22]	77.9	96.0	98.5	64.1	91.0	96.1	523.6
GraDual$^{*}_{\mathrm{WACV'22}}$ [26]	77.0	96.4	98.6	<u>65.3</u>	**91.9**	96.4	525.6
NAAF$^{*}_{\mathrm{CVPR'22}}$ [47]	<u>80.5</u>	<u>96.5</u>	98.8	64.1	90.7	<u>96.5</u>	527.2
CMSEI* (ours)	**81.4**	**96.6**	<u>98.8</u>	**65.8**	<u>91.8</u>	**96.8**	**531.1**
Full 5K							
VSE++$_{\mathrm{BMVC'18}}$ [9]	41.3	69.2	81.2	30.3	59.1	72.4	353.5
SCAN$^{*}_{\mathrm{ECCV'18}}$ [20]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN$^{*}_{\mathrm{ICCV'19}}$ [21]	53.0	81.1	89.4	40.5	70.6	81.1	415.7
IMRAM$^{*}_{\mathrm{CVPR'20}}$ [3]	53.7	83.2	91.0	39.7	69.1	79.8	416.5
CAAN$_{\mathrm{CVPR'020}}$ [48]	52.5	83.3	90.9	41.2	70.3	82.9	421.1
CAMERA$^{*}_{\mathrm{ACMMM'20}}$ [30]	55.1	82.9	91.2	40.5	71.7	82.5	423.9
VSE$\infty_{\mathrm{CVPR'21}}$ [4]	58.3	85.3	<u>92.3</u>	42.4	72.7	<u>83.2</u>	434.3
DIME$_{\mathrm{SIGIR'21}}$ [31]	<u>59.3</u>	<u>85.4</u>	91.9	<u>43.1</u>	<u>73.0</u>	83.1	<u>435.8</u>
CGMN$^{*}_{\mathrm{TOMM'22}}$ [5]	53.4	81.3	89.6	41.2	71.9	82.4	419.8
VSRN++$^{*}_{\mathrm{TPAMI'22}}$ [22]	54.7	82.9	90.9	42.0	72.2	82.7	425.4
NAAF$^{*}_{\mathrm{CVPR'22}}$ [47]	58.9	85.2	92.0	42.5	70.9	81.4	430.9
CMSEI* (ours)	**61.5**	**86.3**	**92.7**	**44.0**	**73.4**	**83.4**	**441.2**

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

Baseline and state-of-the-arts. We compare our proposed CMSEI with several image-sentence retrieval methods on the MS-COCO and Flickr30K datasets in Table 1 and Table 2, including (1) the global matching methods, *i.e.*, VSE++ [9], SGRAF [8], VSE∞ [4] (the reported version with same object inputs), and (2) the attention-based cross-modal interaction methods, *i.e.*, SCAN* [20], CAAN [48], IMRAM* [3], DIME [31] *etc.*, and (3) the graph-based retrieval methods, *i.e.*, VSRN [21], CGMN [5] and GraDual [26], and (4) latest state-of-the-art methods, *i.e.*, DIME [31], NAAF[47], *etc.* Note that, the ensemble models with "*" are further improved due to the complementarity between multiple models. For fair comparison, we also provide the ensemble results, which are averaged similarity scores of image-text model and text-image model. And the results of each single model are provided in Table 3.

Quantitative comparison on MS-COCO. Table 1 lists the experimental results on two kinds of MS-COCO test sets, 5-folds 1K (at the top) and full 5K (at the bottom). Specifically, compared with the state-of-the-art model NAAF [47] on MS-COCO 1K test set, our CMSEI achieves 0.9% and 1.7% improvements in terms of R@1 on both image and sentence retrieval, respectively. Compared with the best cross-modal interaction method DIME [31], CMSEI achieves 2.6% and 1.0% improvements in terms of R@1 on image and sentence retrieval, respectively. And CMSEI clearly outperforms the methods GraDual [26] and CGMN [5], which also employ graph networks, by 5.5% and 10.4% in terms of *rSum*, respectively.

Furthermore, on the larger image-sentence retrieval test data (MS-COCO Full 5K test set), including 5000 images and 25000 sentences, CMSEI outperforms recent methods with a large gap. Following the common protocol [31, 47], CMSEI achieves 5.3%, 15.8%, and 10.3% improvements in terms of *rSum* compared with the latest state-of-the-arts DIME [31], VSRN++ [22] and NAAF [47], respectively. It clearly demonstrates the powerful effectiveness of the proposed CMSEI model with the huge improvements.

Quantitative comparison on Flickr30K. Quantitative results on Flickr30K 1K test set are shown in Table 2, where the proposed approach CMSEI outperforms all state-of-the-art methods in terms of *rSum*. Though for individual recall metrics, we see variations in performance, however, the

Table 2. Comparisons of experimental results on Flickr30K 1K test set. '*' indicates the performance of an ensemble model.

Method	Sentence Retrieval			Image Retrieval			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN$^{*}_{\text{ECCV'18}}$ [20]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
VSRN$^{*}_{\text{ICCV'19}}$ [21]	71.3	90.6	96.0	54.7	81.8	88.2	482.6
CAAN$_{\text{CVPR'20}}$ [48]	70.1	91.6	97.2	52.8	79.0	87.9	478.6
IMRAM$^{*}_{\text{CVPR'20}}$ [3]	74.1	93.0	96.6	53.9	79.4	87.2	484.2
GSMN$^{*}_{\text{CVPR'20}}$ [25]	76.4	94.3	97.3	57.4	82.3	89.0	496.8
CAMERA$^{*}_{\text{ACMMM'20}}$ [30]	78.0	95.1	97.9	60.3	85.9	91.7	508.9
SHAN $^{*}_{\text{IJCAI'21}}$ [14]	74.6	93.5	96.9	55.3	81.3	88.4	490.0
SGRAF$^{*}_{\text{AAAI'21}}$ [8]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
VSE$\infty_{\text{CVPR'21}}$ [4]	81.7	95.4	97.6	61.4	85.9	91.5	513.5
DIME$^{*}_{\text{SIGIR'21}}$ [31]	81.0	95.9	98.4	63.6	**88.1**	**93.0**	520.0
CGMN$^{*}_{\text{TOMM'22}}$ [5]	77.9	93.8	96.8	59.9	85.1	90.6	504.1
VSRN++$^{*}_{\text{TPAMI'22}}$ [22]	79.2	94.6	97.5	60.6	85.6	91.4	508.9
GraDual$^{*}_{\text{WACV'22}}$ [26]	78.3	96.0	98.0	64.0	86.7	92.0	511.4
NAAF$^{*}_{\text{CVPR'22}}$ [47]	81.9	96.1	98.3	61.0	85.3	90.6	513.2
CMSEI* (ours)	**82.3**	**96.4**	**98.6**	**64.1**	87.3	92.6	**521.3**

Table 3. The results of our single CMSEI model, image-to-sentence (I-T) and sentence-to-image (T-I) versions, on MS-COCO and Flickr30K.

Version	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
MS-COCO 5-folds 1K						
I-T	78.7	95.9	98.3	62.5	90.5	96.1
T-I	78.7	96.2	98.8	63.6	90.9	96.2
Flickr30K						
I-T	78.0	95.2	98.0	59.5	84.2	91.1
T-I	79.0	94.6	97.9	60.4	85.5	90.5

Table 4. Comparison results on cross-dataset generalization from MS-COCO to Flickr30k. † means the results are obtained from their published pre-trained model.

Method	Sentence Retrieval		Image Retrieval	
	R@1	R@10	R@1	R@10
VSE++$_{\text{BMVC'18}}$ [9]	40.5	77.7	28.4	66.6
SCAN$^{*}_{\text{ECCV2018}}$ [20]	49.8	86.0	38.4	74.4
CVSE$_{\text{ECCV'20}}$ [35]	57.8	87.2	44.8	81.1
VSE$\infty^{\dagger}_{\text{CVPR'21}}$ [4]	68.0	93.7	50.0	84.9
DIME$^{*\dagger}_{\text{SIGIR'21}}$ [31]	67.4	94.5	53.7	86.5
CMSEI (ours)	**69.6**	**95.2**	**53.7**	**87.2**

proposed CMSEI shows clear improvements under all most metrics compared with the latest state-of-the-art methods.

Generalization ability for domain adaptation. We further validate the generalization ability of the proposed CMSEI on challenging cross-datasets, which is meaningful for evaluating the cross-modal retrieval performance in real-scenario. Specifically, similar to CVSE [35], we transfer our model trained on MS-COCO to Flickr30K dataset. As shown in Table 4, the proposed CMSEI achieves significantly outperforms the baselines. It reflects that CMSEI has an excellent capability of generalization for cross-dataset image-sentence retrieval.

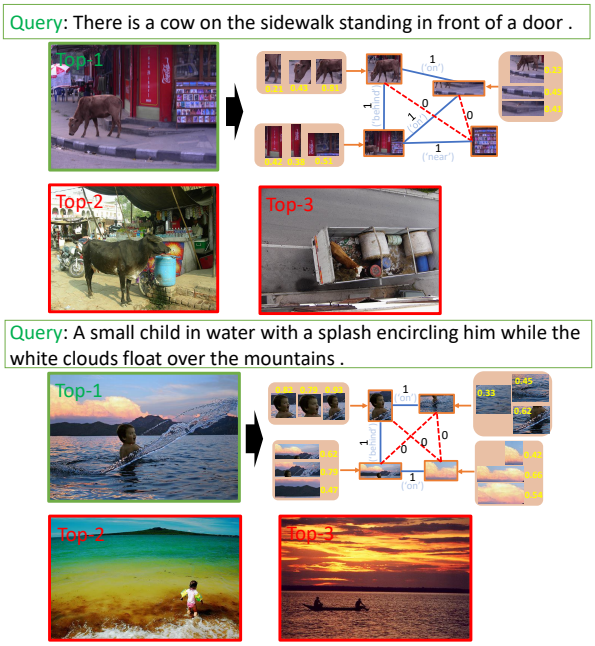

Figure 3. visualization of Top-3 image retrieval results of our CMSEI on MS-COCO (at the top) and Flickr30K (at the bottom). The correctly matched images are marked in green and the mismatched images are marked in red. The learned similarities between objects with high spacial IoUs and the explicit semantic relationship graphs (blue lines mean semantic correlations (indicated by 1), red dashed lines are no significant semantic correlations (indicated by 0)) for the matched image fragments are also partially presented (best viewed in color).

Visualization of results. To better understand the effectiveness of the proposed CMSEI, we visualize matching results from image and sentence retrieval on both MS-COCO and Flickr30K in Figure 3 and Figure 4, respectively. The example on top is from MS-COCO and the one below is from Flick30k. Moreover, we visualize the explicit visual spa-

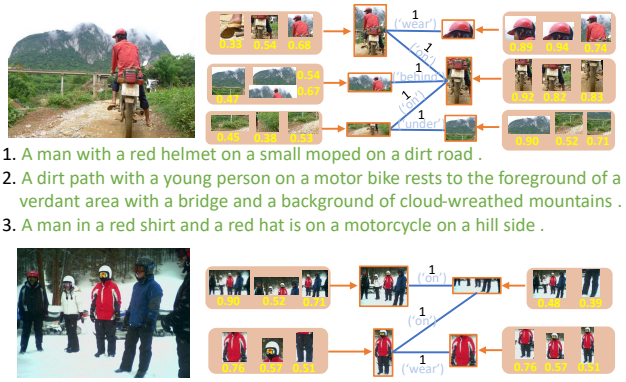

Figure 4. visualization of Top-3 sentence retrieval results of our CMSEI on MS-COCO (at the top) and Flickr30K (at the bottom). The corresponding explicit relationship graphs with the relevant and correlation weights among fragments of images are also partially presented (best viewed in color).

cial and semantic relationship graphs (due to limitations of space we show the graphs partially) for the corresponding images for both retrieval directions, which are used in CMSEI. We also show the learned similarities between objects with high spacial IoUs in space. And the structured correlations among the objects (explicit correlation weight is 1, otherwise 0) can be maintained with the explicit semantic correlation graph guidance. It can be observed that the proposed relationship graphs provide more precise spatial and semantic correlations between the object regions, which can help the model to interact more comprehensively.

4.4. Ablation Studies

We perform detailed ablation studies on Flickr30K to investigate the effectiveness of each component of our proposed CMSEI.

Effects of visual spatial graph. In Table 5, CMSEI decreases absolutely by 2.3% on Flickr30K in terms of $rSum$ when removing the visual spatial graph (w/o VSG). It suggests that the spatial graph reasoning plays an important role in concentrating on spatially relevant regional features for fragments in images. In addition, we achieved slightly lower results using self-attention networks (w. SA) [33], an implicit relationship modeling method, as an alternative to VSG. It demonstrates that our proposed visual spatial graph reasoning can effectively aggregate spatially relevant regional features compared to implicit relational reasoning based on self-attention.

Effects of explicit visual semantic relationship graph. As shown in Table 5, CMSEI decreases absolutely 3.5% in terms of $rSum$ on Flickr30k when replacing explicit visual semantic relationship graph (VSRG) by a fully-connected graph (indicated by w. FG) for images. When dropping VSRG directly (indicated by w/o VSRG), it degrades the $rSum$ by a clear 7.0%. These observations suggest that our explicit VSRG effectively improve visual semantic feature embedding and avoid irrelevant feature incorporation.

Table 5. Ablation studies on Flickr30K 1K test set. All values are ensemble results by averaging two models' (I-T and T-I) similarity.

Method	Sentence Retrieval			Image Retrieval			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
w/o VSG	81.5	95.8	98.6	63.5	87.1	92.5	519.0
w. SA (VSG)	81.2	96.2	98.3	63.5	87.2	92.6	519.0
w/o VSRG	79.5	95.2	98.0	62.3	87.0	92.3	514.8
w. FG (VSRG)	81.0	95.4	98.6	63.0	87.3	92.5	517.8
w/o TG	79.7	95.0	97.9	61.8	86.7	92.2	513.3
w. DTG (TG)	79.6	95.7	97.7	61.6	86.9	92.4	513.9
w/o LLII	73.5	93.6	96.7	57.5	84.2	90.5	496.2
w/o LGII	80.0	95.2	98.2	62.9	87.2	92.3	515.8
CMSEI	**82.3**	**96.4**	**98.6**	**64.1**	**87.3**	**92.6**	**521.3**

Effects of implicit textual graph. When dropping the textual graph reasoning (w/o TG) of sentences, a significant drop in results can be observed. Without implicit semantic reasoning on a fully connected textual graph, we construct a semantic dependency text graph (indicated by w. DTG) by the same way as in method [5], resulting in great degradation. We speculate that these dependencies lose some textual information when interacting across modalities.

Effects of local-local and local-global inter-modal interactions. We evaluate the impact of the local-local and local-global inter-modal interaction (LLII and LGII) for CMSEI. As shown in Table 5, the absence of LLII and the absence of LGII reduce 4.2% and 1.0% in terms of the average of all metrics on Flickr30k, respectively. It is obvious that the multiple inter-modal interactions play a vital role in image-sentence retrieval process, which also suggests that cross-modal interactions effectively narrow the semantic gap between the two modalities.

5. Conclusion

In this paper, we present a cross-modal semantic enhanced interaction method (CMSEI) for image-sentence retrieval. CMSEI engages in (i) enhancing the visual semantic representation with the inter-object relationships and (ii) enhancing the visual and textual semantic representation with multi-level joint semantic correlations on intra-fragment, inter-fragment, and inter-instance. To this end, we propose the intra- and inter-modal semantic correlations and optimize the integrated structured model with cross-modal semantic alignment in an end-to-end representation learning way. Extensive quantitative comparisons demonstrate that our CMSEI achieves state-of-the-art performance on the most of standard evaluation metrics across MS-COCO and Flickr30K benchmarks.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[2] Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. Variational structured semantic inference for diverse image captioning. In *NeurIPS*, pages 1931–1941, 2019.

[3] 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.

[4] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15789–15798, 2021.

[5] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. Cross-modal graph matching network for image-text retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(4):1–23, 2022.

[6] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *NeurIPS*, 28, 2015.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ACL*, 2018.

[8] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, volume 35, pages 1218–1226, 2021.

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[10] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.

[11] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013.

[12] Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *ACM MM*, pages 5185–5193, 2021.

[13] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, pages 6163–6171, 2018.

[14] Zhong Ji, Kexin Chen, and Haoran Wang. Step-wise hierarchical alignment network for image-text matching. *IJCAI*, 2021.

[15] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *ICCV*, pages 5754–5763, 2019.

[16] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[17] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NeurIPS*, pages 1889–1897, 2014.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[19] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *Trans. Assoc. Comput. Linguist*, 2015.

[20] 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.

[21] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4654–4662, 2019.

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[23] 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. Springer, 2014.

[24] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *ACM MM*, pages 3–11, 2019.

[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] Siqu Long, Soyeon Caren Han, Xiaojun Wan, and Josiah Poon. Gradual: Graph-based dual-modal representation for image-text matching. In *WACV*, pages 3459–3468, 2022.

[27] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[28] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017.

[29] Manh-Duy Nguyen, Binh T Nguyen, and Cathal Gurrin. A deep local and global scene-graph matching for image-text retrieval. *arXiv preprint arXiv:2106.02400*, 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 *ACM SIGIR*, pages 1104–1113, 2021.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[34] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *ICLR*, 2016.

[35] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, pages 18–34. Springer, 2020.

[36] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):394–407, 2018.

[37] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016.

[38] 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.

[39] Jonatas Wehrmann, Camila Kolling, and Rodrigo C Barros. Adaptive cross-modal embeddings for image-text alignment. In *AAAI*, volume 34, pages 12313–12320, 2020.

[40] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. Learning dual semantic relations with graph attention for image-text matching. *IEEE Trans. Circuits Syst. Video Technol.*, 31(7):2866–2879, 2020.

[41] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1367–1381, 2017.

[42] 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.

[43] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, pages 4894–4902, 2017.

[44] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. Multi-modal graph contrastive learning for micro-video recommendation. In *ACM SIGIR*, pages 1807–1811, 2022.

[45] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.*, 2:67–78, 2014.

[46] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018.

[47] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *CVPR*, pages 15661–15670, 2022.

[48] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *CVPR*, pages 3536–3545, 2020.

[49] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *CVPR*, pages 10394–10403, 2019.

[50] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(2):1–23, 2020.