

Relation-Aggregated Cross-Graph Correlation Learning for Fine-Grained Image–Text Retrieval

Shu-Juan Peng, Yi He, Xin Liu[✉], *Senior Member, IEEE*, Yiu-ming Cheung[✉], *Fellow, IEEE*,
Xing Xu[✉], *Member, IEEE*, and Zhen Cui[✉], *Member, IEEE*

Abstract—Fine-grained image–text retrieval has been a hot research topic to bridge the vision and languages, and its main challenge is how to learn the semantic correspondence across different modalities. The existing methods mainly focus on learning the global semantic correspondence or intramodal relation correspondence in separate data representations, but which rarely consider the intermodal relation that interactively provide complementary hints for fine-grained semantic correlation learning. To address this issue, we propose a relation-aggregated cross-graph (RACG) model to explicitly learn the fine-grained semantic correspondence by aggregating both intramodal and intermodal relations, which can be well utilized to guide the feature correspondence learning process. More specifically, we first build semantic-embedded graph to explore both fine-grained objects and their relations of different media types, which aim not only to characterize the object appearance in each modality, but also to capture the intrinsic relation information to differentiate intramodal discrepancies. Then, a cross-graph relation encoder is

newly designed to explore the intermodal relation across different modalities, which can mutually boost the cross-modal correlations to learn more precise intermodal dependencies. Besides, the feature reconstruction module and multihead similarity alignment are efficiently leveraged to optimize the node-level semantic correspondence, whereby the relation-aggregated cross-modal embeddings between image and text are discriminatively obtained to benefit various image–text retrieval tasks with high retrieval performance. Extensive experiments evaluated on benchmark datasets quantitatively and qualitatively verify the advantages of the proposed framework for fine-grained image–text retrieval and show its competitive performance with the state of the arts.

Index Terms—Cross-graph relation encoder, fine-grained correspondence, image–text retrieval, intermodal relation.

I. INTRODUCTION

VISION and language are two most prevalent information for human to intuitively understand the real world. With the fast development of multimedia technology, multimedia data, such as image and text, have been accumulated explosively from the social media and web applications. To maximally benefit from the richness of multimedia data, image–text retrieval has become an essential technique for searching engine as well as multimedia data management system, which enables to index semantically relevant instance from one modality with instance from another different modalities. Nevertheless, the modality gap, large intramodal discrepancy, and weak intermodal dependency pose a great challenge to learn the semantic correspondence between the heterogeneous image–text data. For instance, the searching system has to distinguish the phrase “river bank” from “financial bank” (intramodal discrepancies) and connect them to the corresponding visual examples (intermodal dependencies).

In recent years, a great deal of research has been devoted to bridge the heterogeneity gap between image and text, by transforming the heterogeneous data samples into a joint embedding space. Along this line, the pioneer works [1], [2] mainly rely on the handcrafted features extracted from both visual and textual data, which often limit the image–text retrieval performance. Recently, significant progress has been made in representation learning using deep neural networks, and most deep image–text matching methods generally yield the improved retrieval performance on many benchmarks [3], [4]. Early approaches often utilize the global representations to capture the image–text correspondence [5], [6]. Alterna-

Manuscript received 13 September 2021; revised 6 April 2022; accepted 30 June 2022. Date of publication 13 July 2022; date of current version 6 February 2024. This work was supported in part by the Open Research Projects of Zhejiang Lab under Grant 2021KH0AB01, in part by the Fundamental Research Funds for the Central Universities of Huaqiao University under Grant ZQN-709, in part by the National Science Foundation of China under Grant 61673185 and Grant 61976049, in part by the National Science Foundation of Fujian Province under Grant 2020J01083 and Grant 2020J01084, in part by the Natural Science Foundation of Shandong Province under Grant ZR2020LZH008, in part by the National Science Foundation of China (NSFC)/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21, in part by the RGC General Research Fund under Grant 12201321, in part by the Hong Kong Baptist University under Grant RC-FNRA-IG/18-19/SCI/03, and in part by the Innovation and Technology Fund of Innovation and Technology Commission of the Hong Kong Government under Grant ITS/339/18. (*Corresponding author: Xin Liu.*)

Shu-Juan Peng is with the Department of Artificial Intelligence, Huaqiao University, Xiamen 361021, China, also with the Key Laboratory of Pattern Recognition and Computer Vision, Xiamen 361021, China, and also with the Key Laboratory of Computer Vision and Machine Learning (Huaqiao University), Fujian Province University, Xiamen 361021, China (e-mail: pshujuan@hqu.edu.cn).

Yi He is with the Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen 361021, China (e-mail: yhe@hqu.edu.cn).

Xin Liu is with the Department of Computer Science, Huaqiao University, Xiamen 361021, China, and also with the Artificial Intelligence Research Institute, Zhejiang Lab, Hangzhou 311121, China (e-mail: xliu@hqu.edu.cn).

Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China (e-mail: ymc@comp.hkbu.edu.hk).

Xing Xu is with the Center for Future Multimedia and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: xing.xu@uestc.edu.cn).

Zhen Cui is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zhen.cui@njust.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2022.3188569

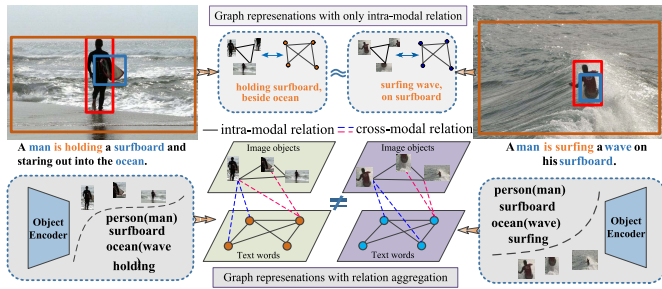


Fig. 1. Illustration of coarse and fine-grained graph correspondence. The existing graph models only aggregate the intramodal relation in each graph representation. Differently, the proposed graph model integrates both of the intramodal relation and intermodal relation to explore the fine-grained relationships across different modalities.

tively, some works map the real-valued feature representations into the Hamming space to improve the retrieval efficiency [7]–[9]. Note that, these approaches generally work well on simple image–text retrieval scenarios that only contain a single object, but which often lead to performance degradation for more realistic cases that involve complex natural scenes.

Recent studies pay attention to local correspondence learning by detecting the fine-grained objects in both images and texts, whereby the image–text similarity scores are aggregated by all salient object pairs for better correlation [10]–[12]. Although these local matching approaches have gained significant improvements over previous global matching works, it remains challenging mainly due to the insufficient representation of object relations and their explicit semantic connections in one modality as it is matched in another modality. For instance, as illustrated in Fig. 1, the main objects “man” and “surfboard” both appear in the scene, but the action meanings between “holding a surfboard” and “on his surfboard” are inherently different. Under such circumstances, the aforementioned methods often fail to explicitly model such semantic relationships, and their performances need further improvements.

In a sense, the exploration of cross-media correlation should not only learn the feature correspondence between the image patches and key words, but also need to characterize the relation correspondence lying in the visual and textual context. In recent years, graph models are popularized to model the objects and their relationships interpretably and have quickly become a powerful tool in high-level image–text matching tasks [13], [14]. Although these graph models have verified the benefits of the graph representations on high-level semantic understanding tasks, the fine-grained cross-modal correspondence may not be fully captured, since the graph representations and correlation regularization are performed in tandem rather than learning simultaneously. Besides, as shown in Fig. 1, these models only consider the intramodal relation in each graph structure, which often ignore the important intermodal relation that can provide complementary information for fine-grained semantic correspondence learning.

To the best of our knowledge, no study has attempted to aggregate both of the intramodal relation and intermodal relation for fine-grained image–text retrieval. Toward this end,

this article presents an efficient relation-aggregated cross-graph (RACG) model to explicitly learn the fine-grained semantic correspondence, by aggregating both intramodal and intermodal relations. On the one hand, the relation-aggregated object correspondence forces the network to explicitly learn the fine-grained semantic correspondence across different modalities. On the other hand, the fine-grained semantic correspondence also promotes to guide the object feature learning process. The main contributions are summarized as follows.

- 1) A cross-graph relation encoder is efficiently addressed to explore the intermodal relationships across different graph representations. This is the first work to perform cross-graph interactions on visual and textual modalities, which can maximally benefit the node-level semantic correlation to infer the fine-grained correspondence.
- 2) A relation-aggregated graph model is newly designed to seamlessly aggregate the intramodal and intermodal relations, which can be explicitly utilized to achieve the fine-grained image–text retrieval.
- 3) Extensive experiments verify the advantages of the proposed framework under various image–text retrieval tasks.

The remaining part of this article is structured as follows. Section II surveys the existing image–text retrieval works, and Section III elaborates the proposed RACG model in detail. The experimental results and quantitative comparisons are provided in Section IV. Finally, we draw a conclusion in Section V.

II. RELATED WORK

Existing image–text retrieval works can be roughly categorized into global correspondence learning and local correspondence learning branches.

Global correspondence learning aims to map the heterogeneous image and text examples into a common embedding space [5], [15]. Along this line, the pioneer canonical correlation analysis (CCA) [1] utilizes the linear transformations to learn a common space that can maximize the correlations between different modalities. Meanwhile, some reasonable extensions, e.g., sparse subspace learning (SSL) [16], [17] and correlated subspace learning (CSL) [2], have also been developed. Note that, these methods highly rely on the handcrafted features extracted from the visual and textual data, which often limit their cross-modal matching performance.

Inspiring from the recent success of deep neural network, some researchers exploit two-branch deep networks to learn the high-level image–text correlations. For instance, Frome *et al.* [18] first proposed a deep visual-semantic matching framework to extract cross-modal representations and then associated them with a structured objective function. Feng *et al.* [19] first employed two uni-modal autoencoders to characterize each modality and then correlated the hidden representations of image–text pairs by minimizing the reconstruction loss. Shu *et al.* [20] and Tang *et al.* [21], respectively, built a deep network structure to translate cross-domain information from text to image, featuring on mitigating the insufficient training data problem. Liu *et al.* [22] utilized a recurrent residual fusion block to correlate the modality-specific representations and created a co-embedding space for

image–text matching. Gu *et al.* [6] selected two generative models to perform similarity matching between textual-visual data samples, while Wang *et al.* [23] utilized a multimodal tensor fusion network to learn the image–text similarity. In addition, Yu *et al.* [24] exploited the potential information of the unlabeled data to contribute the correlation learning among the heterogeneous data. Wang *et al.* [15] investigated two-branch neural networks to learn the similarity between image and text. Alternatively, some works map the deep feature representations into the hash codes to accelerate the retrieval speed [25]–[27]. Note that, these approaches often involve two limitations: 1) these global matching methods often work well on simple image–text retrieval scenario that contains only a single object, but which often degrade their performance on complex natural scenes and 2) the multiple objects and their semantic relations within the data samples are not discriminatively exploited, whereby the semantic correspondence and relationship are not well revealed for high retrieval performance.

Local correspondence learning mainly attempts to learn the local alignment between images and sentences, which have achieved more interpretable retrieval performances [28], [29]. Naturally, a more reasonable way for image–text matching is to capture the fine-grained interplay between the salient image patches and the key words in the sentences. Attention mechanism, aiming at exploiting the salient parts of visual or textual inputs, is recently popularized to learn more discriminative cross-modal representations. For instance, Huang *et al.* [30] addressed a context-modulated attention scheme to correlate a pair of image–text instances, while Nam *et al.* [31] exploited a dual attention network by jointly leveraging visual and textual attentions to estimate the cross-modal similarity. Later, Lee *et al.* [10] presented a stacked cross attention model to discover the full latent alignment between vision and language. In addition, Xu *et al.* [12] proposed a hybrid matching approach that performs cross-modal attention for local semantic alignment. Although these attention mechanisms have achieved impressive image–text retrieval performance, these approaches often lose sight of the relationships between the salient objects in each media data, and therefore, their image–text retrieval performances need further improvements.

With more recent research topics focusing on graph representations, scene graphs are beneficial to model the objects and relationships formally and have quickly become a powerful tool for many high-level semantic understanding tasks. Accordingly, some works attempt different graph structures to represent the visual and textual data. For instance, Li *et al.* [13] built an interpretable reasoning model on a graph topology and performed graph convolutional networks to produce relationship-enhanced features. Liu *et al.* [14] modeled the object, relation, and attribute as a structured phrase and presented a graph structured matching network (GSMN) to learn the semantic correspondence between the structured phrases. Wang *et al.* [32] utilized the graph model to represent the image and text and formulated the image–text retrieval task as a cross-modal scene graph matching (SGM) problem. He *et al.* [33] addressed a cross-graph attention model to guide the feature learning process of each modality and pro-

motored the learning of the shared semantic concepts. Note that, these methods have verified the benefits of the graph representations on high-level semantic understanding tasks. Nevertheless, the current graph models only consider the intramodal relation within each media data representation, which ignore the important intermodal relation that can interactively provide complementary information for semantic correlation learning. From a practical viewpoint, it is still desirable to develop a more robust graph structure for fine-grained image–text retrieval.

III. PROPOSED RACG MODEL

The proposed framework aims to learn the fine-grained correspondence between the image and the text. As illustrated in Fig. 2, we first build semantic-embedded graph to explore the salient patches and their intramodal relations of different media types. Then, a cross-graph relation encoder is efficiently designed to aggregate the intermodal relation across different modalities. Besides, the node-level correspondence module and multihead similarity alignment are leveraged to optimize the graph node representations.

A. Modality-Specific Representation

1) *Visual Feature Embedding*: For an input image \mathbf{I} , the bottom-up attention mechanism [34] is utilized to discriminate a set of objects (i.e., salient image region), with each object represented by a pooled convolutional feature vector, simply denoted as $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_{n_o}\}$, where n_o is the number of detected objects. In general, the position information is able to model the spatial relation of each object [35], [36], and the integration of position information could enhance the discrimination power of the visual representations. Similarly, the absolute normalized position is appended to the corresponding object features, simply written as $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{n_o}\}$. To fully model the relations between image objects, two fully connected layers are, respectively, applied to encode these object features with high-level representation

$$\mathbf{h}_i^{\mathcal{I}} = \text{FC}_o(\mathbf{o}_i) \parallel \text{FC}_p(\mathbf{p}_i) \quad (1)$$

where “ \parallel ” indicates concatenation, and $\text{FC}_o(\cdot)$ and $\text{FC}_p(\cdot)$, respectively, denote the fully connected layer to encode the object vector and position vector. Consequently, the visual representation $\mathbf{H}^{\mathcal{I}} = \{\mathbf{h}_1^{\mathcal{I}}, \mathbf{h}_2^{\mathcal{I}}, \dots, \mathbf{h}_{n_o}^{\mathcal{I}}\}$ can capture both semantic and spatial information to characterize the image sample.

2) *Textual Feature Embedding*: For an input text \mathbf{T} , we first construct the words vocabulary and characterize the i th word with its corresponding index $\mathbf{T}_{\text{word}}^i$. Then, an embedding matrix \mathbf{W}_w is utilized to map the index information into word features. Meanwhile, the off-the-shelf Stanford CoreNLP [37] is utilized to parse the part-of-speech (POS) vector $\mathbf{T}_{\text{pos}}^i$ for the i th word. Similarly, an embedding matrix \mathbf{W}_{pos} is further utilized to encode the POS vector, and we concatenate these two embedding vectors to represent the word

$$\mathbf{t}_i^{\mathcal{T}} = \mathbf{W}_w \mathbf{T}_{\text{word}}^i \parallel \mathbf{W}_{\text{pos}} \mathbf{T}_{\text{pos}}^i \quad (2)$$

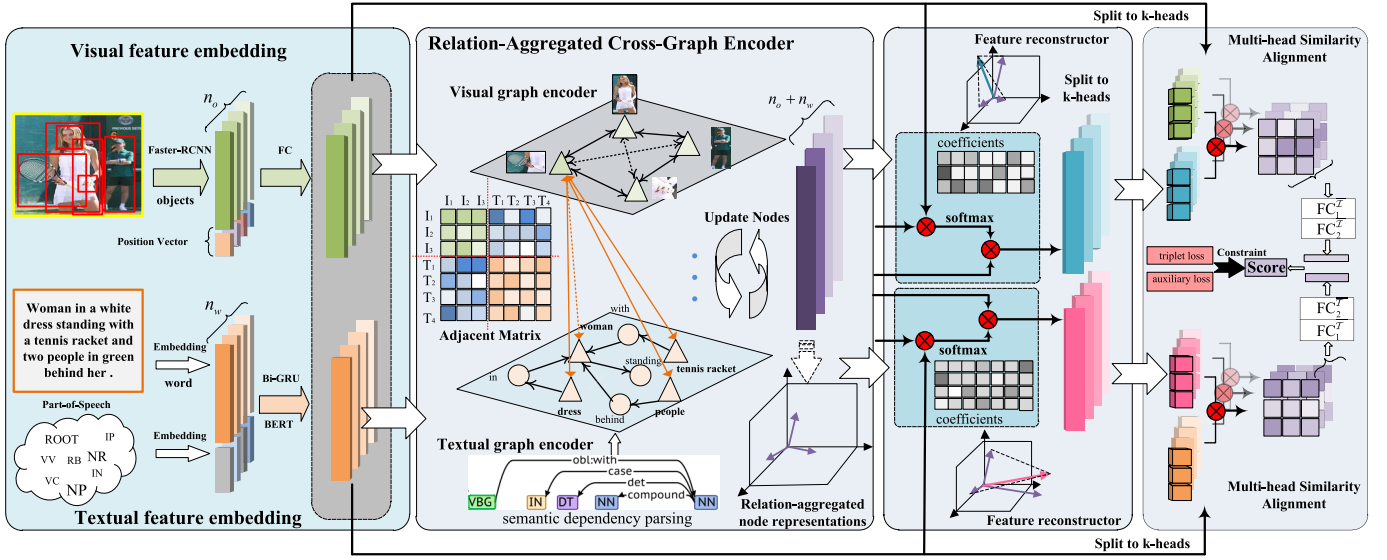


Fig. 2. Schematic architecture of the proposed RACG framework.

where t_i^T encodes the i th word vector that can capture both of the index and POS information in the sentence. Note that, the bidirectional encoding of word information is important for sentence-level representation [34], [38], and we further transform t_i^T into d -dimensional feature space using a bidirectional gated recurrent unit (bi-GRU). That is, the encoding of the i th word is sequentially aggregated by averaging the hidden state of forward and backward bi-GRU

$$\mathbf{h}_i^T = \frac{\overrightarrow{\text{GRU}}(t_i^T) + \overleftarrow{\text{GRU}}(t_i^T)}{2} \quad (3)$$

where $\overrightarrow{\text{GRU}}(t_i^T)$ and $\overleftarrow{\text{GRU}}(t_i^T)$, respectively, denote the aggregated vector of t_i^T derived in forward and backward directions. Consequently, the semantic-enhanced textual representation $\mathbf{H}^T = \{\mathbf{h}_1^T, \mathbf{h}_2^T, \dots, \mathbf{h}_{n_w}^T\}$ is discriminatively obtained to characterize the contextual information in the sentence, where n_w is the number of the word units.

It is noted that the recent Bidirectional Encoder Representations from Transformers (BERT) [39] encoder is another powerful language encoding method, which aims to learn the deep bidirectional representations from the sentence by jointly conditioning on both left and right contexts in all network layers. Alternatively, the BERT encoder is also selected to characterize each word in the sentence

$$\mathbf{h}_i^T = \text{FC}_b(\text{BERT}(t_i^T)) \quad (4)$$

where BERT denotes the bert encoder, and $\text{FC}_b(\cdot)$ denotes the fully connected layer that utilized to map the output of BERT encoder into d -dimensional feature space.

B. RACG Encoder

Inspired by recent advances in graph representation, scene graphs are popularized to model the objects and their relations in high-level semantic understanding tasks. To the best of our knowledge, existing graph models often perform the intramodal relation aggregation and correlation regularization

in a successive way rather than learning simultaneously. Note that, the cross-graph relation, as the complementary, is able to provide valuable intermodal relation for fine-grained semantic correlation learning. To this end, we design an RACG encoder to aggregate both of the intramodal correlations and the intermodal interactions. As shown in Fig. 2, we construct an integrated graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, in which \mathbf{V} denotes a set of nodes that consist of object nodes in image and word nodes in text, and \mathbf{E} is the edge set that represents a connection between two nodes. In particular, the adjacent matrix, usually marked as $\mathbf{A} \in \mathbb{R}^{n \times n}$, is often utilized to represent the relationships between the nodes in \mathbf{V} , with each element A_{ij} characterizing the connection strength between the i th node and the j th node, where $n = n_o + n_w$ is the sum of node numbers. There will be an edge with high weight value connecting two nodes if they have strong semantic relationships, and vice versa. More specifically, the graph encoder \mathcal{G} is comprised of three sub-graph encoders, i.e., visual graph encoder \mathcal{G}_I , textual graph encoder \mathcal{G}_T , and cross-graph encoder \mathcal{G}_C , with architectures detailed as follows.

1) *Visual Graph Encoder*: The scene graph $\mathcal{G}_I = (\mathbf{V}^I, \mathbf{E}^I)$ is constructed from the visual modality, where the salient objects are represented as the nodes $\mathbf{V}^I \in \mathbf{H}^I$ in the graph and are connected by the edge set \mathbf{E}^I . The node relationships are expressed by the weighted adjacent matrix $\mathbf{A}^I \in \mathbb{R}^{n_o \times n_o}$. More formally, an implicit relation between two nodes can be reflected in the form of triplets, e.g., $(\mathbf{h}_i^I, \mathbf{A}_{ij}^I, \mathbf{h}_j^I)$, which semantically describes the semantic connection from head node \mathbf{h}_i^I to tail node \mathbf{h}_j^I . Furthermore, two mapping functions $\kappa(\cdot)$ and $\mu(\cdot)$, respectively, termed head mapping function and tail mapping function, are utilized to map the head node and tail node into the high-level hyperspaces

$$\kappa(\mathbf{h}) = \mathbf{W}_\kappa \mathbf{h} + b_\kappa, \quad \mu(\mathbf{h}) = \mathbf{W}_\mu \mathbf{h} + b_\mu \quad (5)$$

where \mathbf{h} denotes the node feature vector in the graph, and \mathbf{W}_κ , b_κ and \mathbf{W}_μ , b_μ are the trainable parameters. During

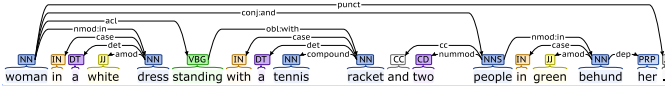


Fig. 3. Dependency parser of a sentence example.

the training process, the mapping functions W_κ and W_μ are initialized by standard Xavier normal initializer, with zero mean and the standard deviation adaptively set at $\sqrt{2/(f_{in} + f_{out})}$, where f_{in} is the input layer size, and f_{out} is the output layer size. Accordingly, the graph self-attention can be utilized to perform node aggregation, and the weighted adjacency matrix can be computed from the hidden representations of each graph node by attending over its neighbors

$$A_{(i,j)}^T = \frac{\exp(\kappa(\mathbf{h}_i^T) \cdot \mu(\mathbf{h}_j^T)^T)}{\sum_{k \in \mathcal{N}_i} \exp(\kappa(\mathbf{h}_i^T) \cdot \mu(\mathbf{h}_k^T)^T)} \quad (6)$$

where $A_{(i,j)}^T$ represents the relation propagation between the i th node and the j th node, and \mathcal{N}_i denotes the neighboring node set of the i th node. In particular, we construct pairwise combinations between all salient objects and utilize the fully connected graph to consider their intramodal relations.

2) *Textual Graph Encoder*: The textual semantic graph $\mathcal{G}_T = (\mathbf{V}^T, \mathbf{E}^T)$ is constructed from the textual modality, where the word units are represented as nodes $\mathbf{V}^T \in \mathcal{H}^T$ in the graph and are connected by the edge \mathbf{E}^T . In particular, the node relationships are implicitly reflected by the weighted adjacent matrix $\mathbf{A}^T \in \mathbb{R}^{n_w \times n_w}$. Similarly, the triplet tuple $(\mathbf{h}_i^T, \mathbf{A}_{ij}^T, \mathbf{h}_j^T)$ is utilized to semantically describe the semantic connection from head node \mathbf{h}_i^T to tail node \mathbf{h}_j^T .

Furthermore, the mapping functions $\kappa(\cdot)$ and $\mu(\cdot)$, formulated in (5), are also utilized to, respectively, map the head node and the tail node into the high-level hyperspaces. That is, the transformed node feature vector $\kappa(\mathbf{h}_i^T)$ is selected as the source of message propagation, while the transformed node feature vector $\mu(\mathbf{h}_j^T)$ is chosen as the destination of message propagation. Similarly, the graph attention is utilized to perform node aggregation, and the weighted adjacency matrix can be computed from the hidden representations of each node in the graph by attending over its neighbors

$$A_{(i,j)}^T = \frac{\exp(\kappa(\mathbf{h}_i^T) \cdot \mu(\mathbf{h}_j^T)^T)}{\sum_{k \in \mathcal{N}_i} \exp(\kappa(\mathbf{h}_i^T) \cdot \mu(\mathbf{h}_k^T)^T)} \quad (7)$$

where $A_{(i,j)}^T$ denotes the relation propagation between the i th node and the j th node in textual graph. As shown in Fig. 3, the off-the-shelf Stanford CoreNLP [37] not only can parse the object (nouns), relation (verbs), and attribute (adjectives or quantifiers) in a sentence, but also encodes their semantic dependencies in the sentences. Accordingly, we set each word as the graph node, and there exists connection between two word nodes if they are semantically dependent.

3) *Cross-Graph Relation Encoder \mathcal{G}_C* : The existing graph models only consider the intramodal relation within each media data representations, which ignore the important

intermodal relation that can interactively provide complementary information for semantic correlation learning. Note that, the node interactions between different modalities are important to benefit the cross-modal correlation learning. Toward this end, a cross-graph relation encoder $\mathcal{G}_C = (\mathbf{V}^C, \mathbf{E}^C)$ is newly designed to bridge the image and text, where the node set consists of pairwise combinations of nodes, respectively, from \mathbf{V}^I and \mathbf{V}^T , i.e., $\mathbf{V}^C = \{(\mathbf{v}_i, \mathbf{v}_j) | \mathbf{v}_i \in \mathbf{V}^I, \mathbf{v}_j \in \mathbf{V}^T\}$. The edge set \mathbf{E}^C assembles the intermodal interactions. As shown in Fig. 2, we group the n_o visual nodes and n_w textual nodes together and construct the cross-modal adjacent matrix \mathbf{A}^C to express the intermodal relation. Accordingly, the triplet tuple $(\mathbf{h}_i^I, \mathbf{A}_{ij}^C, \mathbf{h}_j^T)$ is utilized to semantically characterize the relationship from the head node \mathbf{h}_i^I to the tail node \mathbf{h}_j^T . In this cross-graph encoder, there exist two connection flows, one of which is formed from the image to text, and the other is derived from the text to image. Similarly, the mapping functions $\kappa(\cdot)$ and $\mu(\cdot)$ are further utilized to map the head node and the tail node into the high-level hyperspaces. Then, each visual node associated with its relevant textual nodes or textual node associated with its relevant visual nodes will be aggregated by the cross-graph attention module

$$A_{(i,j)}^{C_{IT}} = \frac{\exp(\kappa(\mathbf{h}_i^I) \cdot \mu(\mathbf{h}_j^T)^T)}{\sum_{k \in \mathcal{N}_i} \exp(\kappa(\mathbf{h}_i^I) \cdot \mu(\mathbf{h}_k^T)^T)} \quad (8)$$

$$A_{(i,j)}^{C_{TI}} = \frac{\exp(\kappa(\mathbf{h}_i^T) \cdot \mu(\mathbf{h}_j^I)^T)}{\sum_{k \in \mathcal{N}_i} \exp(\kappa(\mathbf{h}_i^T) \cdot \mu(\mathbf{h}_k^I)^T)} \quad (9)$$

where $A_{(i,j)}^{C_{IT}}$ denotes the intermodal relation of the i th image node correlating with the j th word node, and $A_{(i,j)}^{C_{TI}}$ represents the intermodal relation of the i th word node correlating with the j th image node. In general, the nouns and attributes within the sentence have a direct semantic connection with the salient objects in the image, while the other words in the sentence contribute little to the cross-modal connection. Therefore, the nouns and attributes within text data are only utilized to form a direct connection with the object nodes in the image.

4) *Relation-Aggregated Integration*: As shown in Fig. 4, the relation-aggregated node representations associated with its relevant nodes from another modality can provide complementary information for learning fine-grained correspondence. Therefore, an efficient graph model should not only aggregate the intramodal relations, but also is capable of assembling the intermodal relations. To this end, we put nodes of two modalities together and build an integrated relation-aggregated graph \mathcal{G} to aggregate both of the intramodal correlations and the intermodal interactions, where its weighted adjacent matrix \mathbf{A} is integrated as follows:

$$\mathbf{A} = \begin{cases} A_{(i,j)}^T/n, & i \leq n_o, \quad j \leq n_o \\ A_{(i,j)}^T/n, & n_o < i \leq n, \quad n_o < j \leq n \\ A_{(i,j)}^{C_{IT}}/n, & n_o < i \leq n, \quad j \leq n_o \\ A_{(i,j)}^{C_{TI}}/n, & i \leq n_o, \quad n_o < j \leq n \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

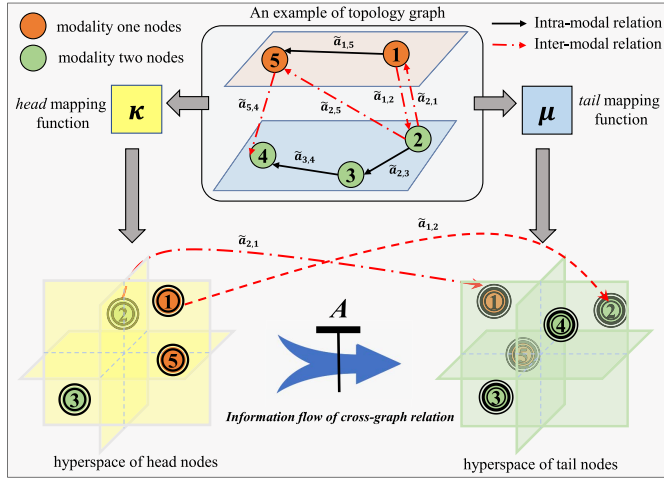


Fig. 4. Illustration of cross-graph relation interaction, in which the circles of different colors denote the graph nodes in different modalities, and the large blue arrow indicates that the relation information in the head hyperspace will flow to the tail hyperspace. The solid line denotes the intramodal relations, while the dashed line of red color represents the intermodal relations.

where $A \in \mathbb{R}^{n \times n}$ denotes the weighted adjacent matrix that expresses an implicit relation aggregated from both of intramodal graph nodes and intermodal graph nodes. Accordingly, the semantic graph correspondence can be explicitly inferred by propagating the neighboring node correspondence, and the representations of graph nodes can be enhanced and updated by aggregating information from their adjacent nodes. To be specific, each node is updated by integrating neighborhood node vectors using graph convolution network (GCN) [32]

$$c_i = \left\| \frac{1}{n} \sum_{k \in \mathcal{N}_i} W_2^c (A_{(i,k)} W_1^c h_k + h_i) + b_c \right\|_2 \quad (11)$$

where c_i is the i th updated node vector, h_i denotes the i th node feature vector in \mathcal{G} , W_1^c , W_2^c , and b_c are trainable parameters, and $\|\cdot\|_2$ denotes the ℓ_2 normalization. Consequently, the relation-aggregated node features $C = \{c_1, \dots, c_{n_o+n_w}\}$ are obtained. Through aggregation and updating of nodes, the dependency between two modalities is well correlated, and the semantic relations across different modalities are enhanced.

C. Node-Level Semantic Correspondence

In essence, each relation-aggregated node should semantically match the input node to maintain the semantic consistency. Therefore, we compute the similarities between the input node and the relation-aggregated nodes and utilize the normalized attention coefficients to indicate their semantic consistency

$$W_{(i,j)}^{\mathcal{I}} = \frac{\exp(\lambda h_i^{\mathcal{I}} (c_j)^T)}{\sum_{k=1}^n (\exp(\lambda h_i^{\mathcal{I}} (c_k)^T))} \quad (12)$$

$$W_{(i,j)}^{\mathcal{T}} = \frac{\exp(\lambda h_i^{\mathcal{T}} (c_j)^T)}{\sum_{k=1}^n (\exp(\lambda h_i^{\mathcal{T}} (c_k)^T))} \quad (13)$$

where $W^{\mathcal{I}} \in \mathbb{R}^{n_o \times n}$ and $W^{\mathcal{T}} \in \mathbb{R}^{n_w \times n}$ are, respectively, the similarity matrix of visual modality and textual modality, and λ is a scaling factor. In general, the node-level correspondence can be indicated by mutual reconstruction from the node-level representations. Accordingly, we reconstruct the nodes as a weighted combination of the relation-aggregated node vectors

$$R^{\mathcal{I}} = W^{\mathcal{I}} C, \quad R^{\mathcal{T}} = W^{\mathcal{T}} C \quad (14)$$

where $R^{\mathcal{I}} = \{r_1^{\mathcal{I}}, \dots, r_{n_o}^{\mathcal{I}}\}$ and $R^{\mathcal{T}} = \{r_1^{\mathcal{T}}, \dots, r_{n_w}^{\mathcal{T}}\}$, respectively, denote the reconstructed node feature vectors of visual and textual modalities. As such, an optimal similarity function, by minimizing the representation learning error for each modality, can be utilized to train the whole model.

D. Multihead Similarity Function

The semantic correspondence can be inferred by computing the similarity score between the node pairs. Unlike previous approaches [13] that compute the global similarity, we employ a multihead module to compute block-wise similarity between the input node and the reconstructed node representation, with large similarity indicating the semantically matched node pair and small similarity indicating the semantically unmatched node pair. To be specific, the multihead mechanism is leveraged to split the i th node feature vector into k -heads

$$h_i^{\mathcal{I}} = h_{i,1}^{\mathcal{I}} \| h_{i,2}^{\mathcal{I}} \| \dots \| h_{i,k}^{\mathcal{I}}, \quad h_j^{\mathcal{T}} = h_{j,1}^{\mathcal{T}} \| h_{j,2}^{\mathcal{T}} \| \dots \| h_{j,k}^{\mathcal{T}} \quad (15)$$

$$r_i^{\mathcal{I}} = r_{i,1}^{\mathcal{I}} \| r_{i,2}^{\mathcal{I}} \| \dots \| r_{i,k}^{\mathcal{I}}, \quad r_j^{\mathcal{T}} = r_{j,1}^{\mathcal{T}} \| r_{j,2}^{\mathcal{T}} \| \dots \| r_{j,k}^{\mathcal{T}} \quad (16)$$

where $\|$ indicates the concatenation operation, $h_{i,k}^{\mathcal{I}}$ is the k th split feature vector from $h_i^{\mathcal{I}}$, and vice versa. Accordingly, we calculate the multiblock similarity for each head

$$s_{i,k}^{\mathcal{I}} = \frac{h_{i,k}^{\mathcal{I}} (r_{i,k}^{\mathcal{I}})^T}{\|h_{i,k}^{\mathcal{I}}\| \|r_{i,k}^{\mathcal{I}}\|}, \quad s_{i,k}^{\mathcal{T}} = \frac{h_{i,k}^{\mathcal{T}} (r_{i,k}^{\mathcal{T}})^T}{\|h_{i,k}^{\mathcal{T}}\| \|r_{i,k}^{\mathcal{T}}\|} \quad (17)$$

where $\|\cdot\|$ denotes ℓ_2 regularization. Accordingly, the output of similarity score is defined as a concatenation over the output of k -heads, followed by two fully connected layers:

$$s_i^{\mathcal{I}} = \text{FC}_2^{\mathcal{I}} (\tanh(\text{FC}_1^{\mathcal{I}} (s_{i,1}^{\mathcal{I}} \| s_{i,2}^{\mathcal{I}} \| \dots \| s_{i,k}^{\mathcal{I}}))) \quad (18)$$

$$s_i^{\mathcal{T}} = \text{FC}_2^{\mathcal{T}} (\tanh(\text{FC}_1^{\mathcal{T}} (s_{i,1}^{\mathcal{T}} \| s_{i,2}^{\mathcal{T}} \| \dots \| s_{i,k}^{\mathcal{T}}))) \quad (19)$$

where $\text{FC}_1^{\mathcal{I}}$, $\text{FC}_2^{\mathcal{I}}$, $\text{FC}_1^{\mathcal{T}}$, and $\text{FC}_2^{\mathcal{T}}$ are trainable parameters, and \tanh denotes the nonlinear activation function. The final similarity S between the image-text pair can be obtained by

$$S = \frac{1}{2} \left(\frac{\sum_{i=1}^{n_o} s_i^{\mathcal{I}}}{n_o} + \frac{\sum_{i=1}^{n_w} s_i^{\mathcal{T}}}{n_w} \right). \quad (20)$$

E. Loss Function

Following most existing works [5], [14], the triplet loss is generally utilized to optimize the hard negative samples. Accordingly, (20) can be employed for regularization at each mini-batch, totally resulting the following loss:

$$\mathcal{L}_{\text{main}} = \sum_{(I,T)} \left([a - S_{IT} + S_{IT}]_+ + [a - S_{IT} + S_{IT}]_+ \right) \quad (21)$$

where $[\cdot]_+ = \max(\cdot, 0)$, and parameter α denotes the margin between the positive pairs and negative pairs. $\hat{\mathcal{I}}$ and $\hat{\mathcal{T}}$ are the corresponding hard negative samples, respectively, obtained by $\hat{\mathcal{I}} = \operatorname{argmax}_{\tilde{\mathcal{I}}} S_{\tilde{\mathcal{I}}\mathcal{T}}$ and $\hat{\mathcal{T}} = \operatorname{argmax}_{\tilde{\mathcal{T}}} S_{\mathcal{I}\tilde{\mathcal{T}}}$, where $\tilde{\mathcal{I}}$ and $\tilde{\mathcal{T}}$ are negative samples. Note that, the triplet loss is able to enlarge the distance between the positive sample pairs and the negative sample pairs, but which cannot ensure that the similarity value between the matched sample pair is large, and the similarity score between the unmatched sample pair is small. To tackle this problem, we further design an auxiliary loss \mathcal{L}_{aux} to increase the absolute score of matched pairs while decreasing the absolute score of unmatched pairs

$$\mathcal{L}_{\text{aux}} = \sum_{(\mathcal{I}, \mathcal{T})} \left([S_{\mathcal{I}\hat{\mathcal{T}}} - \gamma]_+ + [S_{\hat{\mathcal{T}}\mathcal{T}} - \gamma]_+ + [\beta - S_{\mathcal{I}\mathcal{T}}]_+ \right) \quad (22)$$

where β and γ are hyperparameters. Consequently, the total loss is the sum of main loss and auxiliary loss

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \tau \mathcal{L}_{\text{aux}} \quad (23)$$

where hyperparameter τ is utilized to balance the contributions of two loss functions. Through the joint exploitation of (23), the semantic correspondence derived from image and text is well obtained for various image–text retrieval tasks.

F. Testing Stage With ReRanking Strategy

The similarity ranking of the searching results reflects the most relevant results to the user and is of crucial importance to the retrieval systems. Inspired by the reranking scheme proposed by works [11], [23], a reranking process reorganizes the similarity matrix to get a more accurate one, which could narrow the gap between the training and testing data. Similar to work [11], we also select the image-to-text reranking scheme in the experiments. Given a query image \mathbf{I} and its initial ranking list produced by (20), we define $R_{\mathcal{I}\mathcal{T}}(\mathbf{I}, K)$ as the initial cross-modal K -nearest neighbor text of image \mathbf{I}

$$R_{\mathcal{I}\mathcal{T}}(\mathbf{I}, K) = \{\mathbf{T}_1, \dots, \mathbf{T}_j, \dots, \mathbf{T}_K\}. \quad (24)$$

Accordingly, for each candidate text \mathbf{T}_j , a set $R_{\mathcal{I}\mathcal{T}}(\mathbf{T}_j, N)$ of N -nearest images can be defined as

$$R_{\mathcal{I}\mathcal{T}}(\mathbf{T}_j, N) = \{\mathbf{I}_1, \dots, \mathbf{I}_k, \dots, \mathbf{I}_N\} \quad (25)$$

where N is the number of images in the testing set. To fuse the bidirectional nearest neighbors, the position index of each candidate \mathbf{T}_j can be redefined as

$$p(\mathbf{T}_i) = k, \quad \text{if } \mathbf{I}_k = \mathbf{I}, \quad \mathbf{I}_k \in R_{\mathcal{I}\mathcal{T}}(\mathbf{T}_j, N). \quad (26)$$

Then, a position set P can be built for all candidate text in the initial K -nearest neighbors $R_{\mathcal{I}\mathcal{T}}(\mathbf{I}, K)$

$$P(\mathbf{I}, K) = \{p(\mathbf{T}_1), \dots, p(\mathbf{T}_j), \dots, p(\mathbf{T}_K)\}. \quad (27)$$

Consequently, we can refine the retrieval list for the query image \mathbf{I} by reranking the position set as a final retrieval result.

IV. EXPERIMENTS

This section conducts a series of quantitative experiments to validate the efficiency of the proposed framework on fine-grained image–text retrieval task. The performance evaluations and analysis will be detailed in Section IV-A.

A. Dataset and Evaluation Metrics

Two public multimodal datasets, i.e., Flickr30k [40] and Microsoft Common Objects in Context (MSCOCO) [41], are utilized to evaluate the image–text retrieval task, including text retrieval (image query) and image retrieval (text query). Each dataset is briefly described as follows.

- 1) *Flickr30k Dataset*: It consists of 31 783 images collected from the Flickr website, and each image is associated with five sentences. Similar to work [32], we split 1000 images for validation, 1000 images for testing, and the rest for training.
- 2) *MSCOCO Dataset*: It contains 123 287 images, each of which corresponds to five manually annotated sentences. Similar to the work [32], the dataset is divided with 5000 images for validation, 5000 images for testing (MSCOCO 5K), and 113 287 images for training. Meanwhile, we also perform MSCOCO 1K testing in the experiments, in which the test dataset is divided into five 1k subsets, and the image–text retrieval results are the average performance on them.

To quantitatively evaluate the image–text retrieval performance, we report the score of popular $R@K$, which is the percentage of queries whose ground truth is ranked within top- K instances, with higher score indicating the better performance. Meanwhile, we also compute an additional “mR” score for overall evaluation, which averages all the recall values to assess the overall performance for both retrieval tasks.

B. Implementation Details

In the experiments, the proposed framework is implemented in the PyTorch platform. For image representation learning, the pretrained visual features with 36 patches provided by Stacked Cross Attention Network (SCAN) [10] are selected for training, and each patch is characterized with 2048-D vector. Meanwhile, the position information of each region is encoded with 128-D vector. For text representation learning, the word embedding size and the POS size are, respectively, set at 300 and 15. Meanwhile, we utilize either a GRU encoder [34] (word embedding size: 300, batch size: 80) or a pretrained BERT model [39] (word embedding size: 768, batch size: 64) to learn the word-level representations. The output dimension of visual and textual encoder is fixed at 1024. During the training, we utilize Adam optimizer with 25 epochs, and the initial learning rate is set at 0.0002, with decaying 10% every 10 and 15 epochs, respectively, for the Flickr30k and MSCOCO datasets. In multihead similarity function, we set the head number \mathbf{k} at 64 and fix the scaling factor λ at 4. For the regularization parameters, the balance parameter τ is set at 0.2, and the margin values are set at $\alpha = 0.2$, $\beta = 0.7$, and $\gamma = 0.3$.

TABLE I
QUANTITATIVE RESULTS OF IMAGE-TEXT RETRIEVAL ON DIFFERENT DATASETS, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	Backbone (image/text)	Flick30K							MSCOCO 1K								
		Text Retrieval			Image Retrieval				mR	Text Retrieval			Image Retrieval				mR
		R@1	R@5	R@10	R@1	R@5	R@10	R@1		R@5	R@10	R@1	R@5	R@10			
DSPE [3]	VGG-19, w2v+HGLMM	50.1	79.7	89.2	39.6	75.2	86.9	70.1	40.3	68.9	79.9	29.7	60.1	72.1	58.5		
VSE++ [5]	ResNet-152, GRU	52.9	80.5	87.2	39.6	70.1	79.5	68.3	64.6	90.0	95.7	52.0	84.3	92.0	79.8		
TIMAM [4]	ResNet-152, Bert	53.1	78.8	87.6	42.6	71.6	81.9	69.3	-	-	-	-	-	-	-		
GXN [6]	ResNet-152, GRU	56.8	-	89.6	41.5	-	80.1	67.0	68.5	-	97.9	56.6	-	94.5	79.4		
SCO [28]	ResNet-152, LSTM	55.5	82.0	89.3	41.1	70.5	80.1	69.7	69.9	92.9	97.5	56.7	87.5	94.8	83.2		
SCAN* [10]	Faster R-CNN, bi-GRU	67.4	90.3	95.8	48.6	77.7	85.2	77.5	72.7	94.8	98.4	58.8	88.4	94.8	84.7		
CAMP [29]	Faster R-CNN, bi-GRU	68.1	89.7	95.2	51.5	77.1	85.3	77.8	72.3	94.8	98.3	58.5	87.9	95.0	84.5		
CASC* [12]	Faster R-CNN, bi-GRU	68.5	90.6	95.9	50.2	78.3	86.3	78.3	72.3	96.0	99.0	58.9	89.8	96.0	85.3		
PFAN* [35]	Faster R-CNN, bi-GRU	70.0	91.8	95.0	50.4	78.7	86.1	78.7	76.5	96.3	99.0	61.6	89.6	95.2	86.4		
DSRAN* (re-ranking) [11]	Faster R-CNN, Bert	80.5	95.5	97.9	59.2	86.0	91.9	85.2	80.6	96.7	98.7	64.5	90.8	95.8	87.9		
SGM [32]	Faster R-CNN, bi-GRU	71.8	91.7	95.5	53.5	79.6	86.5	79.8	73.4	93.8	97.8	57.5	87.3	94.3	84.0		
VSRN [13]	Faster R-CNN, bi-GRU	71.3	90.6	96.0	54.7	81.8	88.2	80.5	76.2	94.8	98.2	62.8	89.7	95.1	86.1		
GSMN (sparse) [14]	Faster R-CNN, bi-GRU	71.4	92.0	96.1	53.9	79.7	87.1	80.0	76.1	95.6	98.3	60.4	88.7	95.0	85.7		
GSMN (dense) [14]	Faster R-CNN, bi-GRU	72.6	93.5	96.8	53.7	80.0	87.0	80.6	74.7	95.3	98.2	60.3	88.5	94.6	85.3		
GSMN* [14]	Faster R-CNN, bi-GRU	76.4	94.3	97.3	57.4	82.3	89.0	82.8	78.4	96.4	98.6	63.3	90.1	95.7	87.1		
RACG	Faster R-CNN, bi-GRU	76.3	93.2	96.5	57.0	82.1	88.5	82.3	77.6	96.4	98.8	62.2	89.8	95.8	86.8		
RACG (re-ranking)	Faster R-CNN, bi-GRU	78.7	93.9	96.5	57.0	82.1	88.5	82.8	79.7	97.4	98.8	62.2	89.8	95.8	87.2		
RACG	Faster R-CNN, Bert	76.7	94.4	97.6	59.2	85.2	91.5	84.1	78.1	97.2	98.7	63.3	90.2	96.0	87.2		
RACG (re-ranking)	Faster R-CNN, Bert	81.1	96.1	98.0	59.2	85.2	91.5	85.1	81.5	98.0	98.9	63.3	90.2	96.0	88.0		
RACG*	Faster R-CNN, bi-GRU	78.7	94.5	97.9	58.4	83.8	89.6	83.8	78.9	97.5	98.8	64.4	89.7	96.1	87.6		
RACG* (re-ranking)	Faster R-CNN, bi-GRU	80.1	96.5	98.1	58.4	83.8	89.6	84.4	80.0	97.7	98.8	64.4	89.7	96.1	87.8		
RACG*	Faster R-CNN, Bert	81.3	95.5	98.0	60.2	86.6	91.2	85.5	81.1	97.2	98.9	65.7	90.2	96.6	88.3		
RACG* (re-ranking)	Faster R-CNN, Bert	82.2	96.2	98.0	60.2	86.6	91.2	85.7	82.3	97.5	98.8	65.7	90.2	96.6	88.5		

Meanwhile, we compare the proposed RACG model with the state-of-the-art competing methods, including the following: 1) global matching methods: Deep Structure-Preserving Embedding (DSPE) [3], Visual-Semantic Embeddings with hard negatives (VSE++) [5], Text-Image Modality Adversarial Matching (TIMAM) [4], and Generative Cross-modal Network (GXN) [6]; 2) local matching methods: Semantic Concept-Order (SCO) [28], SCAN [10], Cross-modal Adaptive Message Passing (CAMP) [29], Cross-modal Attention with Semantic Consistency (CASC) [12], Position Focused Attention Network (PFAN) [35], and Dual Semantic Relations Attention Network (DSRAN) [11]; and 3) graph matching methods: Visual Semantic Reasoning Network (VSRN) [13], SGM [32], and GSMN [14]. Note that, the best results reported in SCAN [10], GSMN [14], and CASC [12] works are generally obtained by an ensemble of two models or their fused similarities. To be specific, SCAN [10] and PFAN [35] combine the two retrieval models by averaging their predicted similarity scores, while CASC [12] fuses the local attention alignment and global constraint to produce the highest results. GSMN [14] models the text as either a sparse graph or a dense graph and ensembles them by averaging their similarity to improve the performance. Similar to GSMN [14], we also ensemble the similarity of sparse graph (baseline) and dense graph to evaluate the performance. For simplicity, we mark such ensemble method with symbol “*” and report its results.

C. Performance Analysis and Comparison

The image-text retrieval results tested on different datasets are shown in Tables I and II, and it can be found that the global matching methods have delivered relatively lower recall scores, for reason that the semantic correlation between the image and text is not well exploited by global correspondence

learning methods. For instance, GXN [6] utilizes generative models to exploit the global correspondence between the entire image and the sentence, which generally ignores the local structure embedded in real-world data and, therefore, results in a lower matching performance. In contrast to this, the local matching methods often yield the better performances than the global matching methods. For instance, SCAN [10] performs cross-modal attention for local alignment and aggregates the region-word similarity for image-text retrieval, which generally deliver the better image-text matching performances. For example, the $R@1$ scores of text retrieval obtained by the SCAN method are reached up to 67.4 and 72.7, respectively, tested on the Flickr30k and MSCOCO 1K datasets. By considering the relationships between different objects in the scene, the graph matching methods often improve the image-text matching performances. Along this way, the performances delivered by SGM, VSRN, and GSMN methods are generally better than that obtained by global matching methods and most local matching methods. In particular, SGM [32] utilizes the graph representations to characterize both image and sentences, which can well model the salient objects and their high-level semantic relationships. As a result, the $R@1$ scores of text retrieval obtained by the SGM approach reach up to 71.8 and 73.4, respectively, evaluated on the Flickr30k and MSCOCO 1K datasets. This indicates that the high-level semantic relationships can provide valuable information for fine-grained image-text retrieval.

Comparatively speaking, the proposed framework aggregates more semantic relationships by considering both intramodal and intermodal relations, which can explicitly learn the fine-grained semantic correspondence to correlate the image and sentence. As shown in Tables I and II, the proposed RACG framework has yielded comparable and even

TABLE II

QUANTITATIVE COMPARISONS OF IMAGE-TEXT RETRIEVAL ON MSCOCO 5K TEST DATASET, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	MSCOCO 5K test						
	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [5]	41.3	71.1	81.2	30.3	59.4	72.4	68.3
GXN [6]	42.0	-	84.7	31.7	-	74.6	58.3
SCO [28]	42.8	72.3	83.0	33.1	62.9	75.5	61.6
SCAN* [10]	50.4	82.2	90.0	38.6	69.3	80.4	68.5
CAMP [29]	50.1	82.1	89.7	39.0	68.9	80.2	68.3
CASC [12]	47.2	78.3	87.4	34.7	64.8	76.8	64.9
SGM [32]	50.0	79.3	87.9	35.3	64.9	76.5	65.7
VRAN* [13]	53.0	81.1	89.4	40.5	70.6	81.1	69.2
DSRAN*(re-ranking) [11]	57.9	85.3	92.0	40.7	71.2	81.8	72.1
RACG	52.7	81.6	89.8	39.5	70.6	81.0	69.0
RACG (re-ranking)	55.4	84.4	91.0	39.5	70.6	81.0	70.4
RACG (bert)	53.7	82.1	89.9	40.3	70.9	81.3	69.7
RACG (bert, re-ranking)	56.3	84.2	90.7	40.3	70.9	81.3	70.6
RACG*	54.2	81.0	89.5	41.5	71.9	82.1	70.8
RACG* (re-ranking)	57.6	85.6	91.9	41.5	71.9	82.1	71.7
RACG* (bert)	55.3	83.5	90.9	41.7	72.7	82.8	71.2
RACG* (bert, re-ranking)	58.2	86.3	92.0	41.7	72.7	82.8	72.1

better performances than that obtained by other baselines in most cases. It is noted that SGM [32], VSRN [13], and GSMN [14] methods also explore the higher order concepts and their semantic relationships. Nevertheless, the SGM method only considers the intramodal relationships in the graph representation, while the VSRN approach only reasons the relationships of image patches. GSMN [14] attempts to model the objects and relations as the structured phrases and utilizes the graph convolutional layer to learn their semantic correspondences. Similarly, such structured phrase also only considers the intramodal graph representations. In contrast to this, the proposed RACG framework improves the graph representation by extracting both intramodal and intermodal relationships, while considering more discriminative loss function to learn the fine-grained semantic correspondence. As a result, the proposed RACG framework outperforms the global matching methods and most local matching methods by a large margin.

For the Flickr30k dataset, the proposed RACG model yields slightly lower $R@10$ score than that obtained by DSRAN in image retrieval task. In particular, DSRAN develops a multilevel semantic enhancement approach to jointly learn the accurate visual representations, which can, therefore, promote the image retrieval task. By contrast, the proposed RACG model not only achieves the competitive image retrieval performance, but also always delivers the better text retrieval performance. For the MSCOCO 1K dataset, the proposed RACG model delivers a bit lower $R@10$ score and $R@5$ score, respectively, evaluated on text retrieval and image retrieval task. Within these results, the proposed RACG model almost approaches the best scores, and their value margins are very small. Importantly, our proposed model has achieved the best $R@1$ results on all retrieval tasks and simultaneously delivered the best mR values on all datasets. That is, the proposed cross-graph correlation learning scheme is able to index more relevant examples in the ranked one results and is also capable

TABLE III

ABLATION STUDIES TESTED ON Flickr30k DATASET, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

ID	Method	Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
1	w/o image branch	71.5	89.3	94.6	51.6	78.7	86.4
2	w/o text branch	75.1	92.8	95.8	55.6	80.1	86.9
3	w/o cross-graph relation	72.6	91.6	94.6	52.5	79.4	87.4
4	w/o multi-head mechanism	74.3	93.1	95.2	56.8	81.7	87.7
5	w/o image position	76.0	93.1	96.3	56.4	81.9	88.3
6	w/o text part-of-speech	75.8	92.9	96.3	56.8	81.5	87.8
7	w/o auxiliary loss	75.9	93.0	96.0	56.7	82.0	88.0
8	dense graph	75.0	92.7	96.6	55.8	81.1	87.7
9	baseline (sparse graph)	76.3	93.2	96.5	57.0	82.1	88.5

of benefiting the overall cross-modal retrieval performance. For the MSCOCO 5K test set, the proposed RACG model with reranking scheme gains the $R@1$ improvements of 0.3 at text retrieval task and 1.0 at image retrieval task in comparison with the competing DSRAN method. Note that, even if the reranking scheme or ensemble strategy is not enrolled, the proposed RACG model with either BERT encoder or GRU encoder still produces quite competitive performance in comparison with most baselines. This indicates that the proposed framework is capable of indexing much more similar samples in the cross-modal retrieval results.

D. Ablation Studies

To validate the impact of different network modules, we alternatively evaluate the proposed RACG model by attempting different structures: 1) **w/o image branch**: model without the visual-relation aggregation; 2) **w/o text branch**: model without textual-relation aggregation; 3) **w/o cross-graph aggregation**: model without intermodal relation aggregation; 4) **w/o multihead mechanism**: calculate the cosine similarity between the origin features and reconstructed features; 5) **w/o image position**: remove the position information in visual embedding; 6) **w/o text POS**: ignore the POS information in textual embedding; 7) **w/o auxiliary loss**: remove the auxiliary loss; 8) **dense graph**: utilize the dense connections to build textual graphs; and 9) **baseline (sparse graph)**: select the sparse connections to build textual graphs.

The ablation results tested on the Flickr30k dataset are shown in Table III, and it can be found that the removal of visual-relation aggregation significantly degrades the image-text retrieval performance, while the deletion of textual-relation aggregation also hurts the retrieval performances. That is, the graph representations are beneficial to model the objects and relationships efficiently, which can promote the image-text retrieval performance. Meanwhile, the embeddings of position information and the POS information are able to improve the cross-modal retrieval performances. For instance, the embedding of image position gains 0.3% improvement of $R@1$ in text retrieval task and also brings 0.6% improvement of $R@1$ in image retrieval task. That is, the employment of the image position is able to index more relevant examples in the ranked one results and, therefore, benefits the fine-grained cross-modal retrieval performance.







Task	Query	Result (VSRN)	Result (ours)
I→T		<ol style="list-style-type: none"> 2 women are standing in a street playing a blue guitar and a violin A girl is playing the violin in the street while her band mate on the guitar is talking on her cellphone with a confused look . Two women on the street , one is playing the guitar and the other is playing violin . A teenage girl is carrying a guitar in the woods . X A female performer with a violin plays on a street while a woman with a blue guitar looks on . 	<ol style="list-style-type: none"> A female performer with a violin plays on a street while a woman with a blue guitar looks on . 2 women are standing in a street playing a blue guitar and a violin . A female performer with a violin plays on a street while a woman with a blue guitar looks on . Two ladies play the violin and the guitar on the street to entertain the passer byes . Two women on the street , one is playing the guitar and the other is playing violin .
		<ol style="list-style-type: none"> The elderly man is enjoying time in the park visiting with friends. X Several people walking at the park with a little girl in green shirt holding on to an adult wearing a white shirt . X A group of people dancing . X All couples are dancing and enjoying their company . X Men and women sitting and walking around picnic tables and having food . 	<ol style="list-style-type: none"> A group of people sitting at a picnic table eating . Many people are watching street performers dancing . X Men and women sitting and walking around picnic tables and having food . A crowd of people are watch two guys play buckets . X multiple people in a park eating at a picnic table .
T→I	<p>Five snowmobile riders all wearing helmets and goggles line up in a snowy clearing in a forest in front of their snowmobiles; they are all wearing black snow pants and from left to right they are wearing a black coat , white coat, red coat , blue coat , and black coat .</p> <p>A group of people standing in front of an igloo .</p>	 	 

Fig. 5. Visualization of cross-modal retrieval results on the Flickr30k dataset. For each image query, top-5 ranked texts are displayed, and the matched texts are marked as green. For each text query, top-3 ranked images are displayed, and the matched images are highlighted in green.

Furthermore, the performance degradation brought by w/o multihead mechanism indicates that the computation of block-wise similarity can exploit the fine-grained similarity to differentiate the semantically matched and unmatched image-text pairs. Similar to the results in GSMN [14], the sparse textual graph also performs better than the dense graph; this is because the redundant relationships may bring negative impact to the semantically irrelevant words if a fully connected graph is built in textual data.

Comparing with w/o cross-graph aggregation, the proposed RACG model with cross-graph relation aggregation has gained 3.7% and 4.5% improvement of $R@1$ score, respectively, evaluated on the text retrieval and image retrieval tasks. This indicates that the integration of cross-graph relation is capable of providing complementary correlation for fine-grained correspondence learning, which can further strengthen the semantic interaction between different modalities and, therefore, boost the image-text retrieval performances. Remarkably, the proposed RACG approach almost outperforms all of them, for the fact that the aggregation of more informative feature vectors, multihead mechanism, and the cross-graph relation can jointly learn more precise semantic correspondence to promote the image-text retrieval performance.

E. Visualization and Analysis

To verify the superiority of the proposed model, we further visualize some representative image-text retrieval results on the Flickr30k dataset. In particular, we select the Bi-GRU as the text encoder and select the baseline model to index the most relevant examples. Fig. 5 displays top-5 ranked image-to-text (I→T) results and top-3 ranked text-to-image

(T→I) results obtained by the proposed RACG method and the competing VSRN approach. For the first image query, it can be found that the fourth textual result indexed by the VSRN approach fails to match the image query. By contrast, the proposed RACG approach has successfully indexed the semantically relevant textual results. For the second image query, there are many instance objects, i.e., “people” occupy most parts of the image, while the semantics of “eating food” exhibit the weak expression. Under such circumstance, it can be found that the top-4 ranked textual results obtained by VSRN approach are semantically irrelevant to the image query. The main reason lies that the fine-grained relationships among “people,” “eating,” and “picnic table” are not well exploited by the VSRN approach, and the insufficient exploitation of such fine-grained relations may lead to the mismatch. In contrast to this, the proposed RACG framework is able to well learn the fine-grained correspondence of the objects and their potential relations, such as “a group of people” and “eating food,” and three indexed textual results are semantically relevant to the image query. Although two textual examples fail to exactly match the image query, the main semantic descriptions, such as “many people” and “a crowd of people,” are also successfully retrieved. From these results, the proposed RACG method really performs better than the competing VSRN approach. For the first text query, the most relevant image sample indexed by the VSRN approach is ranked at 3, and the rank 1 result is irrelevant. By contrast, the proposed RACG method often retrieves the relevant results with a high rank. That is, the proposed RACG framework is able to precisely learn the fine-grained semantic correspondence between different modalities and, therefore, promotes the retrieval performance.

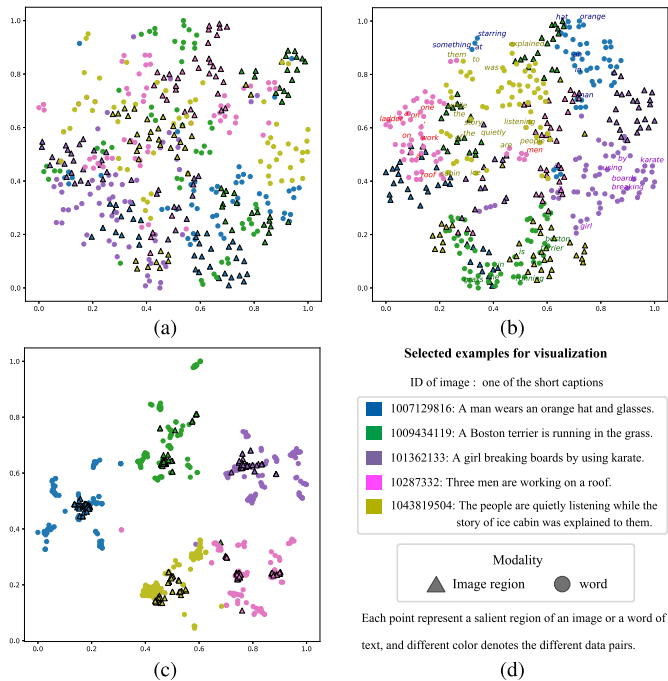


Fig. 6. Visualization of five data pairs from the Flickr30k dataset. (a) Before training process. (b) Intramodal relation aggregation. (c) Relation-aggregated integration. (d) Examples and notations.

Besides, we utilize t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to visualize the learned feature representations, and five examples randomly selected from Flickr30k dataset are chosen for visualization. As shown in Fig. 6(a) and (b), the original feature representations of similar semantics derived directly from the data encoders are scattered far away from each other, and their feature representations aggregated with intramodal relation often gather together. Nevertheless, the intramodal relation aggregation cannot well push the representations of heterogeneous image–text data from the same semantics closer. Comparatively speaking, as shown in Fig. 6(c), the relation-aggregated integration embedded with cross-graph relation not only can push the representations of image–text data from the same semantics closer, but also is able to pull those image–text representations of different semantics away. It indicates that the proposed network structure exhibits high discriminability to learn the semantically differentiable embeddings for each modality, while showing the strong ability to correlate the semantically relevant embeddings from different modalities.

F. Further Analysis

1) *Analysis on Cross-Graph Relation:* The cross-graph relation encoder is reflected in the interactive link between different graph representations, which provides valuable information for fine-grained correspondence learning. As shown in Fig. 7, we visualize some salient regions according to their summed edge weights linking to the neighboring nodes. It can be observed that these salient regions associated with high weights reveal the important semantics of the image, such as “players,” “soccer,” and “green uniforms” in the first picture, and “man,” “hat,” and “glasses” in the second picture.

Meanwhile, some node pairs with relatively high weights in \mathcal{G}_T , e.g., “players-passing-others,” “players-uniform,” “wearing-hat,” and “man-ear-pierced,” often contain very discriminative semantic information to correlate the text and images, and it can also be observed that several node pairs with higher edge weights linking from the image nodes to the text nodes are also semantically relevant to each other. Therefore, the high-level semantic information and cross-graph relation are interacted across the image–text node pairs, which promote to aggregate the fine-grained semantic correspondence.

Besides, we further select the Bert encoder as the baseline and make a quantitative analysis of different text node representations. As shown in Fig. 8, it can be found that the dense connections between regions and words could induce the irrelevant and redundant connection, which degrade the image–text retrieval performance. Meanwhile, the utilization of both noun and adjective nodes often boosts the image–text retrieval performances. That is, the adjective nodes in text enrich the cross-modal interactions and also provide valuable information to correlate the semantically relevant image examples.

2) *Analysis on Auxiliary Loss Function:* To verify the effectiveness of the designed loss functions, we draw the loss curves and monitor the variations of auxiliary loss by adding or dropping it from the integrated loss function. Fig. 9 shows the loss values and retrieval results with the changing of iteration numbers. On the one hand, it can be observed that the whole loss function will converge to a lower value if the auxiliary loss is integrated to learn together. On the other hand, the embedding of auxiliary loss will also contribute to improve the image–text retrieval performances, especially at the early iterations. Therefore, the auxiliary loss has a very positive impact to the image–text retrieval results.

3) *Analysis on Training Time:* To show the time complexity of the proposed framework, we record the execution times at each epoch. The proposed model is trained on GPU NVIDIA RTX 2080Ti, and we select the competing GSMN method for comparison. Since the proposed RACG approach integrates more feature types, graph modules and loss functions to discriminatively learn the relation-aggregated graph representations, the execution time of training time or testing time could be much higher than that obtained by the competing GSMN method. Fortunately, as illustrated in Table IV, the proposed RACG method does not significantly increase the training time and testing time to a large extent, while achieving the best retrieval performances. From a practical viewpoint, the proposed RACG method achieves a good balance between the time cost and retrieval performance, which is suitable for fine-grained image–text retrieval tasks.

4) *Analysis on Model Parameters:* The multihead mechanism is utilized to compute block-wise similarity between the node pairs, which can jointly exploit the semantic correspondence at different feature positions. Within the multihead mechanism, we further explore the impact of head number k in (15) and (16) and set the number value ranging from 8 to 128. Representative results tested on the Flickr30k and MSCOCO datasets are shown in Fig. 10, and it can be found that the different settings of the head number could affect the

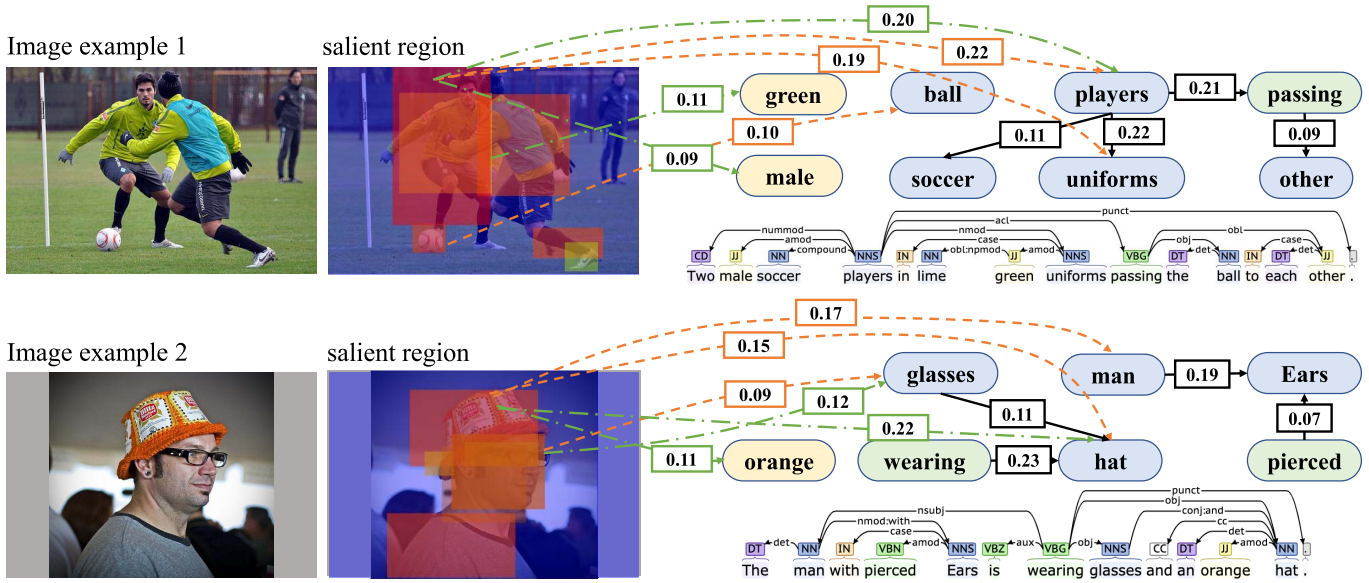


Fig. 7. Cross-graph relation visualization. The first column shows the image examples, and the second column displays the top-3 ranked salient regions from \mathcal{G}_V , and top-3 ranked regions linking from the text to the image in \mathcal{G} . These salient regions are visualized with different colors according to their edge weights, and the warmer red indicates that the region aggregates more relation information with other regions. In the third column, we quantify top-4 ranked edges (black arrow) in \mathcal{G}_T . The orange dashed arrow indicates that the nouns are only utilized to connect image nodes, while the green dashed arrow indicates that the nouns and adjectives are enrolled to connect image nodes.

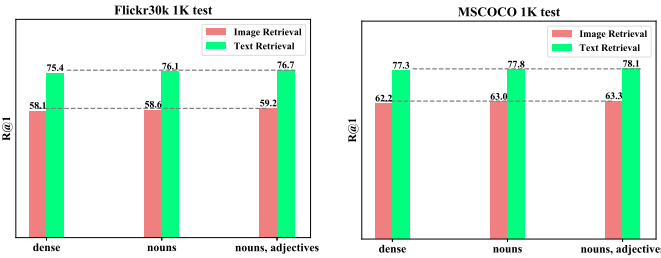


Fig. 8. Quantitative analysis of different graph connections.

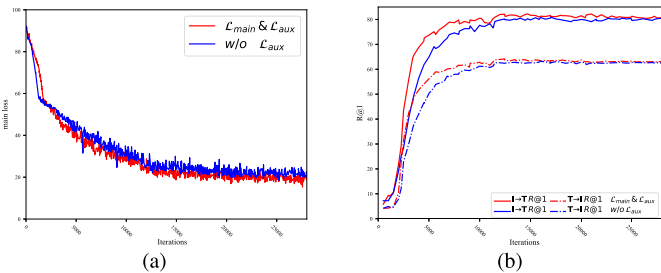


Fig. 9. Illustration of auxiliary loss on image-text retrieval results. (a) Loss convergence curve. (b) Retrieval result in terms of $R@1$.

image-text matching accuracy to some degree, but not in a large magnitude. Remarkably, the proposed RACG model delivers the best results when the head number k is equal to 64. On the one hand, if the head number is too small, the block-wise similarity in multihead mechanism is not well exploited, such that the derived similarity is not discriminative for fine-grained semantic matching. On the other hand, if the head number is too large, the dimension of the block features will be very small, making it insufficient for precisely expressing

TABLE IV
EVALUATION OF EXECUTION TIMES ON Flickr30k DATASET

Method	Training (h/epoch)	Testing (s)	mR (%)
GSMN (sparse)	0.41	134.3	80.0
w/o image branch	0.40	164.4	78.7
w/o text branch	0.52	174.7	81.1
w/o cross graph attention	0.58	155.2	79.7
w/o multi-head mechanism	0.65	175.0	81.5
w/o image position	0.64	183.3	82.0
w/o text part-of-speech	0.63	182.5	81.8
w/o auxiliary loss	0.63	183.4	81.9
baseline (sparse Graph)	0.66	183.4	82.3

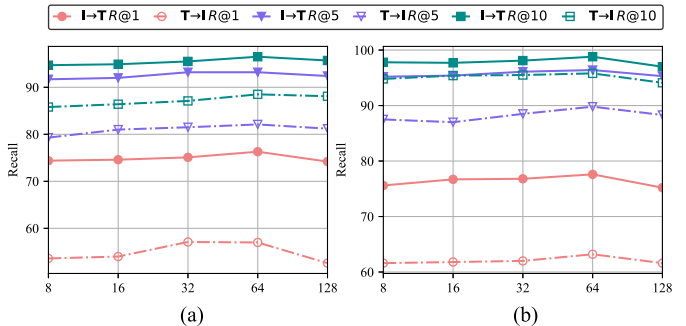


Fig. 10. Impacts of head number k in multihead mechanism. (a) Num of heads on Flickr30k. (b) Num of heads on MSCOCO.

the object features in images and words in sentences. Similar to most works [12], [13], the different value settings of other balance parameters only induce a minor fluctuation to the retrieval performance, and these parameters are generally insensitive to the image-text retrieval performances within a

wide range of values. Experimentally, the suggested values always deliver the competing performances.

V. CONCLUSION

In this article, we propose an efficient RACG model to achieve fine-grained image–text retrieval. Within the proposed framework, a relation-aggregated graph model is newly designed to explore the fine-grained relationships across different modalities, which can mutually boost cross-modal interactions to learn more precise intermodal dependencies. Accordingly, the representations of each node on the newly designed graph model are optimized by aggregating both intramodal and intermodal relations. Meanwhile, the feature reconstruction module and multihead similarity are seamlessly integrated to jointly optimize the node-level semantic correspondence, whereby the derived feature embeddings aggregated in such graph model are discriminatively obtained to benefit the fine-grained image–text retrieval in a more interpretable and plausible way. Extensive experiments conducted on various kinds of image–text retrieval tasks have shown its outstanding performance.

Along the line of this work, several open problems also deserve our further research. For example, the current graph model attempts to enhance the representations of each node by aggregating both of the intramodal relations and the intermodal relations. If the object number in an image and the word number in the sentence are very large, the integrated graph model will be very huge, and the updating of graph nodes will need more computational load. Therefore, it is necessary to study a fast graph node aggregation method for processing the complex data samples. In addition, the salient object detection methods would also have an influence on the fine-grained image–text matching results, and more robust salient object detection methods deserve further studies.

REFERENCES

- [1] J. C. Pereira *et al.*, “On the role of correlation and abstraction in cross-modal multimedia retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.
- [2] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, “Joint feature selection and subspace learning for cross-modal retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.
- [3] L. Wang, Y. Li, and S. Lazebnik, “Learning deep structure-preserving image-text embeddings,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.
- [4] N. Sarafianos, X. Xu, and I. Kakadiaris, “Adversarial representation learning for text-to-image matching,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5814–5824.
- [5] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “VSE++: Improving visual-semantic embeddings with hard negatives,” in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–14.
- [6] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, “Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.
- [7] Z. Ji, Y. Sun, Y. Yu, Y. Pang, and J. Han, “Attribute-guided network for cross-modal zero-shot hashing,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 321–330, Jan. 2020.
- [8] X. Lu, L. Liu, L. Nie, X. Chang, and H. Zhang, “Semantic-driven interpretable deep multi-modal hashing for large-scale multimedia retrieval,” *IEEE Trans. Multimedia*, vol. 23, pp. 4541–4554, 2021.
- [9] X. Liu, X. Wang, and Y.-M. Cheung, “FDDH: Fast discriminative discrete hashing for large-scale cross-modal retrieval,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 12, 2021, doi: 10.1109/TNNLS.2021.3076684.
- [10] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.
- [11] K. Wen, X. Gu, and Q. Cheng, “Learning dual semantic relations with graph attention for image-text matching,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2866–2879, Jul. 2021.
- [12] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, “Cross-modal attention with semantic consistence for image-text matching,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5412–5425, Dec. 2020.
- [13] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4653–4661.
- [14] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, “Graph structured network for image-text matching,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10921–10930.
- [15] L. Wang, Y. Li, J. Huang, and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.
- [16] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, “Sparse multi-modal hashing,” *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.
- [17] X. Zhai, Y. Peng, and J. Xiao, “Learning cross-media joint representation with sparse and semisupervised regularization,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.
- [18] A. Frome *et al.*, “DeViSE: A deep visual-semantic embedding model,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.
- [19] F. Feng, X. Wang, and R. Li, “Cross-modal retrieval with correspondence autoencoder,” in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 7–16.
- [20] X. Shu, G.-J. Qi, J. Tang, and J. Wang, “Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation,” in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 35–44.
- [21] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang, “Generalized deep transfer networks for knowledge propagation in heterogeneous domains,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 4s, pp. 1–22, Nov. 2016.
- [22] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, “Learning a recurrent residual fusion network for multimodal matching,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4127–4136.
- [23] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, “Matching images and text with multi-modal tensor fusion and reranking,” in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 12–20.
- [24] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, “Adaptive semi-supervised feature selection for cross-modal retrieval,” *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1276–1288, May 2019.
- [25] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, “Instance-aware hashing for multi-label image retrieval,” *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2469–2479, Jun. 2016.
- [26] L. Jin, X. Shu, K. Li, Z. Li, G.-J. Qi, and J. Tang, “Deep ordinal hashing with spatial attention,” *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2173–2186, May 2019.
- [27] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, “MTFH: A matrix trifactORIZATION hashing framework for efficient cross-modal retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 964–981, Mar. 2021.
- [28] Y. Huang, Q. Wu, C. Song, and L. Wang, “Learning semantic concepts and order for image and sentence matching,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6163–6171.
- [29] Z. Wang *et al.*, “CAMP: Cross-modal adaptive message passing for text-image retrieval,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5764–5773.
- [30] Y. Huang, W. Wang, and L. Wang, “Instance-aware image and sentence matching with selective multimodal LSTM,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7254–7262.
- [31] H. Nam, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2156–2164.
- [32] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, “Cross-modal scene graph matching for relationship-aware image-text retrieval,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1497–1506.
- [33] Y. He, X. Liu, Y.-M. Cheung, S.-J. Peng, J. Yi, and W. Fan, “Cross-graph attention enhanced multi-modal correlation learning for fine-grained image-text retrieval,” in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1865–1869.

- [34] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
- [35] Y. Wang *et al.*, "Position focused attention network for image-text matching," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3792–3798.
- [36] L. Qu, M. Liu, D. Cao, L. Nie, and Q. Tian, "Context-aware multi-view summarization network for image-text matching," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1047–1055.
- [37] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.
- [38] T.-J. Fu, P.-H. Li, and W.-Y. Ma, "GraphRel: Modeling text as relational graphs for joint entity and relation extraction," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1409–1418.
- [39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [40] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Feb. 2014.
- [41] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.



Shu-Juan Peng received the Ph.D. degree in computer science from Wuhan University, Wuhan, China, in 2009.

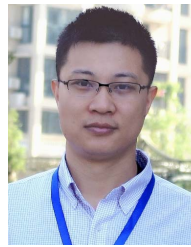
She is currently an Associate Professor with the Department of Artificial Intelligence, Huaqiao University, Xiamen, China, and also a Research Fellow with the Key Laboratory of Pattern Recognition and Computer Vision, Xiamen, and the Key Laboratory of Computer Vision and Machine Learning (Huaqiao University), Fujian Province University, Xiamen. Her research interests include multimedia

data analysis, pattern recognition, and computer animation.



Yi He received the M.S. degree in software engineering from Huaqiao University, Xiamen, China, in 2022.

He is currently a Research Fellow with the Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Xiamen, and the Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen. His current research interests include multimedia data analysis, pattern recognition, and deep learning.



Xin Liu (Senior Member, IEEE) received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, SAR, China, in 2013.

He was a Visiting Scholar with the Computer and Information Sciences Department, Temple University, Philadelphia, PA, USA, from 2017 to 2018. He is currently a Full Professor with the Department of Computer Science and Technology, Huaqiao University, Xiamen, China, and also a Research Fellow with the Zhejiang Lab, Hangzhou, China. His current research interests include multimedia data analysis and deep learning.



Yiu-ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, in 2000.

He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China. His current research interests include machine learning, pattern recognition, and visual computing.

Prof. Cheung is a fellow of the Institution of Engineering and Technology (IET) and British Computer Society (BCS). He is the Founding Chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He served as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2014 to 2020, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, and *Pattern Recognition*.



Xing Xu (Member, IEEE) received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively, and the Ph.D. degree from Kyushu University, Fukuoka, Japan, in 2015.

He is currently with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include information retrieval, pattern recognition, and computer vision.



Zhen Cui (Member, IEEE) received the B.S. degree from Shandong Normal University, Jinan, China, in 2004, the M.S. degree from Sun Yat-sen University, Guangzhou, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences, Beijing, China, in 2014.

He was a Research Assistant with Nanyang Technological University (NTU), Singapore, in 2012. He was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore, from 2014 to 2015. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. His research interests include deep learning, computer vision, and pattern recognition.