Learning Relationship-Enhanced Semantic Graph for Fine-Grained Image–Text Matching

Xin Liu, *Senior Member, IEEE*, Yi He, Yiu-Ming Cheung, *Fellow, IEEE*, Xing Xu, *Member, IEEE*, and Nannan Wang, *Member, IEEE*

***Abstract*—Image–text matching of natural scenes has been a popular research topic in both computer vision and natural language processing communities. Recently, fine-grained image–text matching has shown its significant advance in inferring the high-level semantic correspondence by aggregating pairwise region–word similarity, but it remains challenging mainly due to insufficient representation of high-order semantic concepts and their explicit connections in one modality as its matched in another modality. To tackle this issue, we propose a relationship-enhanced semantic graph (ReSG) model, which can improve the image–text representations by learning their locally discriminative semantic concepts and then organizing their relationships in a contextual order. To be specific, two tailored graph encoders, visual relationship-enhanced graph (VReG) and textual relationship-enhanced graph (TReG), are respectively exploited to encode the high-level semantic concepts of corresponding instances and their semantic relationships. Meanwhile, the representations of each graph node are optimized by aggregating semantically contextual information to enhance the node-level semantic correspondence. Further, the hard-negative triplet ranking loss, center hinge loss, and positive–negative margin loss are jointly leveraged to learn the fine-grained correspondence between the ReSG representations of image and text, whereby the discriminative cross-modal embeddings can be explicitly obtained to benefit various image–text matching tasks in a more interpretable way. Extensive experiments verify the advantages of the proposed fine-grained graph matching approach, by achieving the state-of-the-art image–text matching results on public benchmark datasets.**

***Index Terms*—Contextual information, high-level semantic concept, image–text matching, relationship-enhanced graph.**

Manuscript received 8 April 2022; accepted 26 May 2022. Date of publication 20 June 2022; date of current version 17 January 2024. This work was supported in part by the Open Project of Zhejiang Lab under Grant 2021KH0AB01 and Grant 2021KG0AB01; in part by the National Science Foundation of China under Grant 61673185, Grant 61672444, Grant 61976049, Grant 61922066, and Grant 61876142; in part by the NSFC/RGC Joint Research Scheme under Grant N_HKBU214/21; in part by the General Research Fund of Research Grants Council (RGC) under Grant RGC/HKBU/12201321; in part by Hong Kong Baptist University under Grant RC-FNRA-IG/18-19/SCI/03 and Grant RC-IRCMs/18-19/SCI/01; in part by the Innovation and Technology Fund of Innovation and Technology Commission of the Government of the Hong Kong SAR under Grant ITS/339/18; in part by the National Science Foundation of Fujian Province under Grant 2020J01084; and in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15. This article was recommended by Associate Editor P. Shi. *(Corresponding author: Yiu-Ming Cheung.)*

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).

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

Yiu-Ming Cheung is with the Department of Computer Science and Institute of Research and Continuing Education, 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 610051, China (e-mail: xing.xu@uestc.edu.cn).

Nannan Wang is with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: nnwang@xidian.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2022.3179020.

Digital Object Identifier 10.1109/TCYB.2022.3179020

I. Introduction

With the fast development of multimedia technology, multimedia data, such as image and text, has been emerging rapidly and accumulated explosively on the Internet. In order to maximally benefit from the richness of multimedia data, image–text matching has become an essential technique for searching engine as well as multimedia data management system, featuring on providing flexible retrieval experience to index semantically relevant instance from one modality to another modality. In recent years, image–text matching has attracted considerable attention in multimedia community, and such technique has been widely applied for various applications, such as image–sentence matching [1], cross-modal retrieval [2], image captioning [3], visual question answering [4], and so forth. In this work, we mainly focus on the problem of the image–text matching, which aims to measure the similarity between images and textual sentences, for example, given an image query to find similar sentences, namely, image-to-text matching, and given a sentence query to retrieve semantically matched images, called text-to-image retrieval.

The key challenge of image–text matching lies in correctly understanding their semantic concepts, discover their full latent semantic correspondence and measuring their semantic similarity. In recent years, a great deal of research has been devoted to bridge the semantic gap between image and textual sentences, either in learning global correspondence [5], [6] or local correspondence [7]. The global correspondence learning methods aim to jointly project the entire image and text data into a common latent space for heterogeneity minimization, whereby the mapping features of image and text in such latent space can be directly measured. Remarkably, these approaches attempt global representations to express the whole image and sentence, which ignore the importance of local cross-modal similarities and are therefore limited on the simple image–text matching scenario that contains only a single object. Since the semantically relevant data of different modalities often

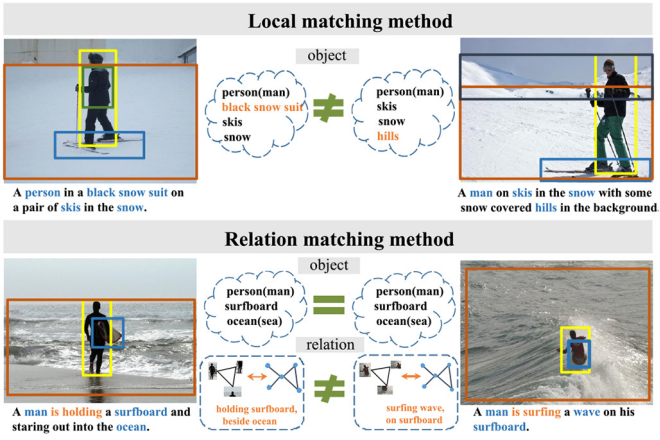

Fig. 1. Illustration of different semantic matching mechanisms.

have an unequal amount of information, such globally semantic matching approaches often degrade their performance for more realistic cases that involve complex natural scenes.

Local correspondence learning, aiming to capture the fine-grained interplay between image and text data, is another more interpretable branch of image–text matching way [8]–[11]. Along this line, salient image patches are detected and image–text similarity scores are aggregated by all or salient region–word pairs, which have gained significant improvements over previous global correspondence learning works. As illustrated in Fig. 1, the upper pictures and captions are quite similar, but which show the different scenes. For instance, the main objects *person*, *black snow suit*, *skis*, and *snow* are appeared in the top-left scene, while the object *black snow suit* disappears in the top-right scene and *hill* appears. Under such circumstances, the local correspondence learning methods are able to distinguish these differences. Although a lot of progress has been developed in this learning area, it is still a challenge problem mainly due to the complex visual semantic discrepancy. More specifically, the aforementioned methods often ignore the relationships between different fine-grained patches. In practice, a natural scene contains not only several objects but also exhibits their interactions, relative positions, and high-level semantic relationships, which are equally important to the image–text matching problem.

Inspired by recent advances in graph representation, scene graphs are popularized to model the objects and their relationships formally and have quickly become an efficient tool in high-level semantic understanding tasks [12], [13]. Although these graph matching methods have shown the impressive image–text matching performance, the derived semantic concepts within these models are generally tangled with each other and the importance of different relationships are not fully exploited. It is noted that different organizations of semantic concepts may lead to completely diverse semantic meanings, which would have different contributions for image–text matching. As shown in Fig. 1, the bottom scenes almost share the similar objects and captions, but the relative positions between the detected objects are slightly different, for example, *holding* specified in the left scene and *surfing* revealed in the right scene. Therefore, it is necessary to pay more attention to some informative relationships for discriminative analysis. In addition, sentence descriptions are weak annotations, where the words in a sentence correspond to some particular, but unknown regions in the image. For fine-grained image–text matching, the efficient extraction of the meaningful words and their semantic relationships is imperative to further guide the fine-grained object correspondence learning.

Motivated by the rapid success of graph representation that can flexibly learn high-level semantic concepts and their relationships, we propose a relationship-enhanced semantic graph (ReSG) model, which can enhance the image–text representations to boost fine-grained cross-modal matching. To be specific, two tailored graph models, visual relationship-enhanced graph (VReG) and textual relationship-enhanced graph (TReG), are respectively exploited to encode the high-level semantic concepts of corresponding instances and their semantic relationships contextually. The VReG encoder enhances the representations of each node on the visual graph by aggregating useful concept information from other nodes and weighting their contextual relationships, while the TReG encoder exploits a bidirectional semantic graph to jointly encode the forward and backward relationship information. As a result, the relationship-enhanced semantic features can be well aggregated in each graph, and the image–text matching problem can be converted into finding the similarity between these two tailored graphs. The proposed approach improves the state-of-the-art methods by providing the following four contributions.

1) A relationship-enhanced graph matching network is explicitly developed to improve the image–text representation, with each graph node aggregating semantically contextual information to learn fine-grained correspondence.
2) A bidirectional textual graph encoder is discriminatively proposed to jointly encode the object-level and relationship-level features for text data while embedding the forward and backward contextual information.
3) The center hinge loss and positive–negative margin loss are introduced to guide the fine-grained object correspondence and relationship correspondence learning.
4) Extensive experiments verify the advantages of the proposed approach under various image–text matching scenarios and show its superiority over the state of the arts.

The remainder of this article is structured as follows. Section II surveys the existing multiview and cross-modal anomaly detection works, and Section III elaborates the proposed relationship-enhanced graph model in detail. The experimental results are provided in Section IV. Finally, we draw a conclusion in Section V.

II. Related Work

The key issue of image–text matching is to measure the semantic similarity between visual and textual inputs, and various kinds of matching works have been developed, either in global matching or local matching ways. This section briefly surveys the representative methods of these two aspects.

A. Global Matching Methods

Global matching methods mainly learn the semantic correspondence between the whole image and sentence. Along this line, canonical correlation analysis (CCA) [14] is possibly the most popular baseline for image–text matching, which aims to find linear projections that maximize the correlation between the projected vectors from different modalities. Later, many improved extensions, for example, latent subspace analysis (LSA) [15] and correlated subspace learning (CSL) [16], have also been developed. It is noted that these methods often limit their capacities for processing large-scale and high-dimensional multimodal data [17]. Alternatively, cross-modal hashing, aiming to transform the high-dimensional data into compact binary codes [18], [19], has also been developed to reduce storage cost and speed up the retrieval speed. Nevertheless, these methods inevitably impose additional binary constraints during the hash code learning process, which may accumulate large quantization error to degrade the matching performance.

With the recent advances of deep learning in multimedia applications, deep neural network (DNN) is popularized to extract powerful visual and textual features. Along this way, Andrew *et al.* [20] exploited a deep CCA structure to maximize the correlation between image and text data. Feng *et al.* [21] addressed a correspondence autoencoder (Corr-AE) to correlate hidden representations of image and text modalities. Huang *et al.* [22] presented a selective multimodal long short-term memory network (sm-LSTM) for instance-aware image and sentence matching. Huang *et al.* [11] exploited a modal-adversarial hybrid transfer network (MHTN) to handle the insufficient training data within cross-modal retrieval tasks. Gu *et al.* [23] incorporated generative processes into the image–text embedding learning process, while Wehrmann and Barros [24] designed an efficient character-level inception module to learn textual semantic embeddings. Similarly, Faghri *et al.* [6] exploited the hardest negative samples to learn the multimodal embeddings, while Wang *et al.* [25] investigated two-branch neural networks to learn an explicit shared latent embedding space. Differently, Peng and Qi [26] addressed a reinforced cross-media bidirectional translation approach to model the correlation between visual and textual descriptions. It is noted that the primary drawback of these methods is that they generally consider the correspondence between the entire image and sentence [27], [28], which often ignore their fine-grained correspondence. As a result, their cross-modal matching performances are often not satisfactory for more realistic cases that involve complex natural scenes.

B. Local Matching Methods

The local matching methods primarily explore the local alignment between image regions and sentence words, for reason that the words in a sentence correspond to some particular regions in an image. Accordingly, Karpathy and Fei-Fei [8] leveraged multimodal RNN to detect local image regions and align them with words in the sentence. Huang *et al.* [11] utilized a multiregional CNN to predict the semantic concepts and employed a conventional LSTM to perform image–sentence matching. Wang *et al.* [7] considered the fine-grained cross-modal interactions and designed cross-modal adaptive message passing (CAMP) to filter out irrelevant information. Besides, some works focus on salient regions and words by using an attention mechanism. For instance, Nam *et al.* [9] proposed dual attention networks (DANs) to attend salient regions in images and specific words in text data. Similarly, Lee *et al.* [10] employed the stacked cross attention to attend salient regions and key words. Wei *et al.* [29] presented a multimodality cross attention (MMCA) network for image and sentence matching by jointly modeling the intramodality and intermodality relationships in a unified deep model. Within these approaches, the image–text similarities are aggregated by salient region–word pairs with various kinds of attention mechanisms. Nevertheless, these methods often ignore the fine-grained relationships within the salient regions or words, and their local correspondences are not sufficiently exploited for better cross-modal retrieval performance.

In recent years, some approaches attempt multilevel information to learn more precise image–text correspondence. For instance, Ma *et al.* [1] jointly mapped global image–text pair, local regions, and words into a common space and implicitly learned the region–word correspondence. Wu *et al.* [30] exploited the local and global semantic consistencies to guide the learning of common embeddings. Qi *et al.* [31] presented the cross-media relation attention network to explore global, local, and relation alignments across different media types. Xu *et al.* [32] jointly utilized cross-modal attention for local alignment and multilabel prediction for global semantic consistency. Peng *et al.* [33] proposed a multilevel adaptive alignment approach to explore global, local, and relation alignments. Wang *et al.* [34] exploited the consensus information by computing the statistical co-occurrence correlations between the semantic concepts from the image captioning corpus and deploying the constructed concept correlation to yield the consensus-aware concept representations. It is noted that these approaches still lack of mining the object relationships between different local patches, which could provide rich complementary hints for fine-grained correlation learning.

With more recent research topics focusing on the objects and relationships in the scene [35]–[37], graph models are introduced to model the objects and relationships interpretably. For instance, Yuan *et al.* [36] employed the cross-modal graph to mine the intermodality relationships among the RGB-D scenes. Liu *et al.* [37] proposed a language-guided graph representation to capture the grounding entities and their relations and developed a cross-modal graph matching strategy for multiple-phrase visual grounding task. For image–text matching, Wang *et al.* [12] utilized the graph model to model the image and text and jointly characterized the objects and relationships for the efficient image–text retrieval. Liu *et al.* [13] modeled the object, relation, and attribute as a structured phrase and designed a graph-structured matching network (GSMN) to learn fine-grained image–text correspondence. In a sense, the semantic relationships between textual words are very weak, and these graph models do not sufficiently consider the bidirectional contextual relationships in a sentence and the

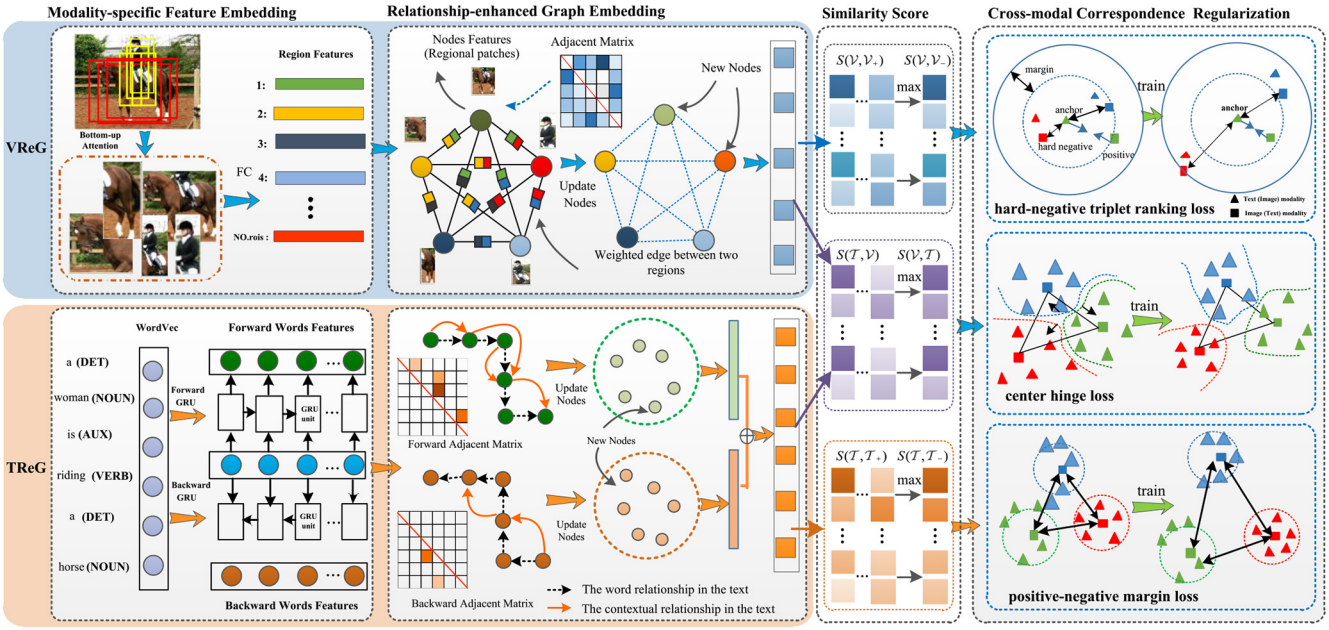

Fig. 2. Schematic architecture of the proposed framework with ReSG encoders.

derived graph models are not discriminative enough. Different from existing approaches, the proposed framework aims to discriminatively discover the latent relationships between the salient image patches and key words, while aggregating more semantically contextual information for inferring fine-grained image–text similarity.

III. Proposed Method

The overall framework of the proposed model is illustrated in Fig. 2, which consists of two tailored graph encoders, that is, VReG and TReG. The VReG encoder enhances the representations of each node on the visual graph by aggregating useful concept information from other nodes and weighting their contextual relationships, while the TReG encoder exploits a bidirectional semantic graph to jointly encode the forward and backward relationship information. These two models are trained correlatively by an efficient loss function, and the derived relationship-enhanced graph model can be well utilized to perform image–text matching. This section shall first elaborate these two graph encoders in tandem and then detail the loss function for fine-grained correspondence learning.

A. Visual Relationship-Enhanced Graph

1) Visual Feature Embedding: For fine-grained image analysis, it is imperative and efficient to detect the salient image patches, while depicting their relationships. Similar to work [10], we utilize ResNet-101 [38] as the basic network and employ bottom-up-attention [39] to detect the instances and salient objects in an image. Accordingly, the category of instance and the object attribute can be well obtained. Experimentally, the pretraining region features provided from work [10] are selected for fast training, and we denote the set of object features as $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{n_o}\}, \mathbf{C} \in \mathbb{R}^{n_0 \times 2048}$, where n_o is the number of detected objects in an image. Further, a fully connect layer is then applied to transform these object features to a d-dimensional embedding space

$$\mathbf{H}^v = \mathbf{W}_l \mathbf{C} + \mathbf{b}_l \tag{1}$$

where $\{\mathbf{W}_l, \mathbf{b}_l\}$ are the trainable parameters of the fully connect layer. Accordingly, a group of visual object features $\mathbf{H}^v = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n_o}\}, \mathbf{H}^v \in \mathbb{R}^{n_0 \times d}$ can be obtained, where $\mathbf{h}_i$ is the ith feature vector of the detected object in an image.

2) Contextual Visual Graph Encoder: As shown in Fig. 1, the left-bottom image is expressed as "a man holding a surfboard and staring out into the sea," while the right-bottom image is displayed as "a man is surfing on a surfboard." It can be found that the significant objects, such as "man," "surfboard," and "sea," are detected in both images. If we only utilize the region features to express the image, these two images may be judged as similar ones. Apparently, the relationship of "holding" or "surfing" is the key difference between these two scenes. In general, the object detection often extracts a fixed number of salient regions, and some images may not have such number of significant regions. Therefore, the detected image patches may exist overlap between regional expressions, and we can assume that there are potential connections between the significant regions. To this end, we design an undirected weighted relationship graph $\mathbf{G}^v = (\mathbf{V}^v, \mathbf{E}^v)$, where node $\mathbf{V}^v \in \mathbf{H}^v$ is the set of object features and edge $\mathbf{E}^v$ represents the relationship between two connected nodes and regularized by weighted adjacent matrix $\mathbf{A}^v \in \mathbb{R}^{n_0 \times n_0}$. Note that the weights of edges indicate the degree of relationship between two connected nodes, and the weighted edge $\mathbf{e}_{ij}^v$ between the ith object and the jth object is computed by

$$\mathbf{e}_{ij}^v = \mathrm{ReLU}\Big(\big(\mathbf{W}_1^v \mathbf{h}_i^v + \mathbf{b}_1^v\big)\big(\mathbf{W}_2^v \mathbf{h}_j^v + \mathbf{b}_2^v\big)\Big) \tag{2}$$

where $\{\mathbf{W}_1^v, \mathbf{b}_1^v, \mathbf{W}_2^v, \mathbf{b}_2^v\}$ are the trainable parameters to calculate the weights of graph edges. Accordingly, the weighted

adjacent matrix can be defined as

$$\mathbf{A}^v_{(i,j)} = \begin{cases} \mathbf{e}^v_{ij}, & \text{if } i \neq j \\ 0, & \text{else}. \end{cases} \tag{3}$$

Accordingly, the contextual visual graph $\mathbf{G}^v$ can be well constructed, which can well utilized to model the relationship associations between different visual objects.

3) Visual Relationship-Enhanced Aggregator: The visual graph is able to model the semantic relationship between different image objects, and the image objects can aggregate significant information that related to each other by using the weighted edges. To this end, the graph nodes are aggregated with other semantically correlated objects by

$$\mathbf{v}_i = \text{ReLU}\left(\sum_{j=1}^{n_o} \mathbf{A}^v_{(i,j)} \times \left(\mathbf{W}^v_c \mathbf{h}^v_j + \mathbf{b}^v_c\right) + \mathbf{h}^v_i\right) \tag{4}$$

where $\{\mathbf{W}^v_c, \mathbf{b}^v_c\}$ are the trainable convolution parameters, $\mathbf{v}_i \in \mathbb{R}^{1\times D}$ is the ith node in visual graph. Further, these updated visual features $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n_o}\}$ are further normalized with ℓ_2 norm: $\mathcal{V} = \|\mathbf{V}\|_2$, and the relationship-enhanced visual features are aggregated in the final graph representation $\mathcal{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n_o}\} \in \mathbb{R}^{n_0 \times D}$.

B. Textual Relationship-Enhanced Graph

Textual relationship mining is a task of understanding the fine-grained parts of a natural language sentence. In general, the text data is a sequence with tagging information, and the bidirectional analysis of text data is also of crucial importance to the discriminative analysis. For instance, the meanings of sentences "a man is surfing the wave on his surfboard" and "a man is surfing down a hill of sand" are different. If we read the sentences phrase by phrase from front to back, the word "man" will first appear, and followed by the word "surfing" in the mind. Under such circumstances, the pictures described by these two captions might be the same in your mind. Differently, if we read the sentences word by word from back to front, "wave on his surfboard" or "a hill of sand" will first appear in mind, which directly shows the different scenarios. Therefore, we argue that the high-level text feature representations derived from forward and backward understanding of sentence are more discriminative to characterize the word-level semantic information.

1) Bidirectional Textual Feature Embedding: In a sense, sentence descriptions are weak annotations, which make it difficult to guide the fine-grained object correspondence learning. Differing from other methods that directly embed the index of the word vocabulary as word features [40], we enhance the representation by incorporating the part-of-speech information into text, and take the word as the smallest unit to extract textual information in a sentence. Formally, given a text $\mathbf{T}$, we first utilize one-hot vector $\mathbf{I}^i_{\text{word}}$ to represent the index position of the ith word in the entire vocabulary, and then map this word into a 300 dims vector through the embedding matrix $\mathbf{W}_{e\_\text{word}}$. Then, we select the spacy tool to detect the part-of-speech vector $\mathbf{I}^i_{\text{pos}}$ of the ith word, and map this vector into a 15 dims vector through the embedding matrix $\mathbf{W}_{e\_\text{pos}}$. Accordingly, we concatenate these two embedding vectors as the word representation vector

$$\mathbf{w}_i = \text{Concat}\left(\mathbf{W}_{e\_\text{word}}\mathbf{I}^i_{\text{word}}, \mathbf{W}_{e\_\text{pos}}\mathbf{I}^i_{\text{pos}}\right) \tag{5}$$

where $\mathbf{w}_i \in \mathbb{R}^{1\times 315}$ is the ith word vector that contains the index and part-of-speech information in the sentence. As discussed in Section II, the object relationships between different words are important for high-level semantic understanding tasks. Differing from exiting methods that directly employ Bi-GRU to encode the text into global features [7], [41], we construct two relation graphs by, respectively, aggregating forward contextual information and backward contextual information. To be specific, the feature derived from forward GRU is sequentially aggregated from the first word until the last word of the sentence

$$\overrightarrow{\mathbf{h}}^t_i = \overrightarrow{\text{GRU}}(\mathbf{w}_i), \ \overrightarrow{\mathbf{h}}^t_i \in \mathbb{R}^{1\times d_a} \tag{6}$$

where $\overrightarrow{\mathbf{h}}^i_t$ is the ith forward hidden state that aggregated the word feature in forward direction, and d_a is the aggregated feature dimension. In backward GRU, the feature representation of words is aggregated in reverse order from the last word to the first word of the sentence

$$\overleftarrow{\mathbf{h}}^t_i = \overleftarrow{\text{GRU}}(\mathbf{w}_i), \quad \overleftarrow{\mathbf{h}}^t_i \in \mathbb{R}^{1\times d_a} \tag{7}$$

where $\overleftarrow{\mathbf{h}}^t_i$ is the ith backward hidden state that aggregated the word feature in backward direction. Accordingly, two kinds of contextual feature vectors $\overrightarrow{\mathbf{H}} = \{\overrightarrow{\mathbf{h}}^t_1, \overrightarrow{\mathbf{h}}^t_2, \ldots, \overrightarrow{\mathbf{h}}^t_{n_w}\} \in \mathbb{R}^{n_w\times d_a}$ and $\overleftarrow{\mathbf{H}} = \{\overleftarrow{\mathbf{h}}^t_1, \overleftarrow{\mathbf{h}}^t_2, \ldots, \overleftarrow{\mathbf{h}}^t_{n_w}\} \in \mathbb{R}^{n_w\times d_a}$ are discriminatively obtained to characterize each word, respectively, memorizing all word vectors in forward and backward orders.

2) Bidirectional Textual Graph Encoder: The Bi-GRU network performs well in memorizing the neighboring word information in a sentence, but which lacks of interaction between key words that are far from each other. For instance, as shown in Fig. 2, the backbone words of textual data are "woman," "riding," and "horse," and these keywords are not adjacent. Therefore, the features derived only from Bi-GRU network cannot aggregate all the relationships and interactions between these keywords. Specifically, the graph structure is also appropriate for characterizing the relationships in a sentence [40], and we further organize the word features of the input sentence into a contextual graph. To this end, we build up a bidirectional weighted relation graph to enhance the word representation, while considering the semantic correlation between the keywords in a contextual semantic order. Accordingly, two directed weighted graphs $\mathbf{G}^t_f = (\mathbf{V}^t_f, \mathbf{E}^t_f)$ and $\mathbf{G}^t_b = (\mathbf{V}^t_b, \mathbf{E}^t_b)$ are exploited to, respectively, characterize the forward relationship and backward relationship in the textual sequence, where $\mathbf{V}^t_f$ (or $\mathbf{V}^t_b$) is the set of word features $\overrightarrow{\mathbf{H}}^t$ (or $\overleftarrow{\mathbf{H}}^t$) and edge set $\mathbf{E}^t_f$ (or $\mathbf{E}^t_b$) is described by weighted adjacent matrix $\mathbf{A}^t_f$ (or $\mathbf{A}^t_b$). To construct the textual relationship, we calculate the cosine distance between word features in $\overrightarrow{\mathbf{H}}^t$ and $\overleftarrow{\mathbf{H}}^t$. Note that different word relationships would have different contributions to characterize the sentence and some of

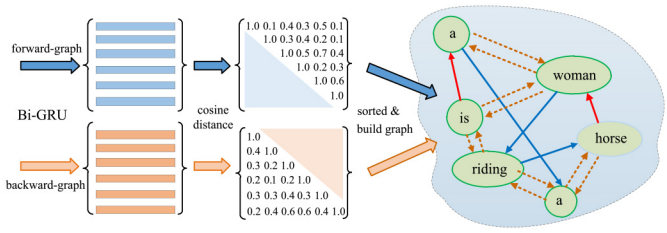

Fig. 3. Overview of constructing bidirectional textual graphs.

them are even redundant. For the ith word vector $\overrightarrow{\mathbf{h}}_i^t$ (or $\overleftarrow{\mathbf{h}}_i^t$), its relationship with other word vectors will be ranked according to their cosine distances. The edge exists between the ith word and the jth word if their cosine distance is ranked within top k neighbors. As depicted in Fig. 3, we calculate the forward weight $\overrightarrow{\mathbf{e}}_{ij}^t$ and backward weight $\overleftarrow{\mathbf{e}}_{ij}^t$ to represent the graph edge, respectively, via the following equations:

$$\overrightarrow{\mathbf{e}}_{ij}^t = \mathrm{ReLU}\left(\left(\overrightarrow{\mathbf{W}}_1^t \overrightarrow{\mathbf{h}}_i^t + \overrightarrow{\mathbf{b}}_1^t\right)\left(\overrightarrow{\mathbf{W}}_2^t \overrightarrow{\mathbf{h}}_j^t + \overrightarrow{\mathbf{b}}_2^t\right)\right) \tag{8}$$

$$\overleftarrow{\mathbf{e}}_{ij}^t = \mathrm{ReLU}\left(\left(\overleftarrow{\mathbf{W}}_1^t \overleftarrow{\mathbf{h}}_i^t + \overleftarrow{\mathbf{b}}_1^t\right)\left(\overleftarrow{\mathbf{W}}_2^t \overleftarrow{\mathbf{h}}_t^j + \overleftarrow{\mathbf{b}}_2^t\right)\right) \tag{9}$$

where $\{\overrightarrow{\mathbf{W}}_1^t, \overrightarrow{\mathbf{b}}_1^t, \overrightarrow{\mathbf{W}}_2^t, \overrightarrow{\mathbf{b}}_2^t\}$ and $\{\overleftarrow{\mathbf{W}}_1^t, \overleftarrow{\mathbf{b}}_1^t, \overleftarrow{\mathbf{W}}_2^t, \overleftarrow{\mathbf{b}}_2^t\}$ are, respectively, the trainable parameters to calculate the weighted edge of forward and backward graphs. Then, the weighted adjacency matrices of the forward relation graph and backward relation graph can be, respectively, defined as

$$\mathbf{A}_{(i,j)}^f = \begin{cases} \overrightarrow{\mathbf{e}}_{ij}^t, & \text{if } \overrightarrow{\mathbf{h}}_j^t \in \mathbf{N}_k\left(\overrightarrow{\mathbf{h}}_i^t\right), \quad i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$\mathbf{A}_{(i,j)}^b = \begin{cases} \overleftarrow{\mathbf{e}}_{ij}^t, & \text{if } \overleftarrow{\mathbf{h}}_j^t \in \mathbf{N}_k\left(\overleftarrow{\mathbf{h}}_i^t\right), \quad i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $\mathbf{N}_k(\cdot)$ is the top-k nearest neighbor set that ranked by the cosine distances, $\mathbf{A}^f \in \mathbb{R}^{n_w \times n_w}$ and $\mathbf{A}^b \in \mathbb{R}^{n_w \times n_w}$ are, respectively, the weighted adjacency matrices of corresponding forward relation graph and backward relation graph.

3) Textual Relationship-Enhanced Aggregator: Within the constructed graphs, the bidirectional word features are updated and aggregated with word affinities in the sentence. In forward relationship-enhanced graph or backward relationship-enhanced graph, the edge and weight, respectively, represent the semantic association and their degree between two words. Therefore, other words are attended to the current word by the weighted edges, and the aggregated word features are semantically correlated with other words. Accordingly, the updating vectors are formulated by

$$\overrightarrow{\mathbf{t}}_i = \mathrm{ReLU}\left(\sum_{j=1}^{n_w} \mathbf{A}_f^{(i,j)} \times \left(\overrightarrow{\mathbf{W}}_f^t \overrightarrow{\mathbf{h}}_j^t + \overrightarrow{\mathbf{b}}^t\right)\right) + \overrightarrow{\mathbf{h}}_i^t \tag{12}$$

$$\overleftarrow{\mathbf{t}}_i = \mathrm{ReLU}\left(\sum_{j=1}^{n_w} \mathbf{A}_b^{(i,j)} \times \left(\overleftarrow{\mathbf{W}}_b^t \overleftarrow{\mathbf{h}}_j^t + \overleftarrow{\mathbf{b}}^t\right)\right) + \overleftarrow{\mathbf{h}}_i^t \tag{13}$$

where $\{\overrightarrow{\mathbf{W}}_f^t, \overrightarrow{\mathbf{b}}^t\}$ and $\{\overleftarrow{\mathbf{W}}_b^t, \overleftarrow{\mathbf{b}}^t\}$ are trainable convolution parameters, respectively, for forward and backward relation graphs. Consequently, the updated forward word feature vector $\mathbf{T}_f = \{\overrightarrow{\mathbf{t}}_1, \overrightarrow{\mathbf{t}}_2, \ldots, \overrightarrow{\mathbf{t}}_{n_w}\} \in \mathbb{R}^{n_w \times D}$ and backward word feature vector $\mathbf{T}_b = \{\overleftarrow{\mathbf{t}}_1, \overleftarrow{\mathbf{t}}_2, \ldots, \overleftarrow{\mathbf{t}}_{n_w}\} \in \mathbb{R}^{n_w \times D}$ are well obtained. Finally, we further average these updated word vectors and normalize them with ℓ_2 norm

$$\mathcal{T} = \left\| \frac{\mathbf{T}_f + \mathbf{T}_b}{2} \right\|_2 \tag{14}$$

where $\mathcal{T} = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_{n_w}\} \in \mathbb{R}^{n_w \times D}$ is the final bidirectional relationship-enhanced word feature vectors.

C. Similarity Function

In the ReSG model, the visual graph nodes $\mathcal{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n_o}\}$ and textual graph nodes $\mathcal{T} = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_{n_w}\}$ are aggregated with the relationship-enhanced semantic features, each of which is a D-dimension vector. Inspired by work [12], we utilize the cosine distance to measure the similarity between two node vectors $\boldsymbol{v}_i$ and $\boldsymbol{\tau}_j$ as $\boldsymbol{\tau}_j^{\mathrm{T}} \boldsymbol{v}_i$. Accordingly, we calculate the similarity scores of all visual and textual graph nodes, and then get an $n_w \times n_o$ score matrix. That is, each node in visual and textual graphs will match with nodes from another modality graph to learn the node correspondence. For the cross-modal matching problem, we first find the maximum value of each row to pick up the most related visual object (or textual word), and then sum these matched values to express the overall similarity score $S(\mathcal{V}, \mathcal{T})$ [or $S(\mathcal{T}, \mathcal{V})$]

$$S(\mathcal{V}, \mathcal{T}) = \sum_{j=1}^{n_w} \max_{i \in [1, n_o]} \left(\boldsymbol{\tau}_j^{\mathrm{T}} \boldsymbol{v}_i\right) \tag{15}$$

$$S(\mathcal{T}, \mathcal{V}) = \sum_{i=1}^{n_o} \max_{j \in [1, n_w]} \left(\boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{\tau}_j\right). \tag{16}$$

D. Loss Function

The triplet loss is often utilized to regularize the correspondence between image and text, which forces the similarity score of the matched image–text pairs to be generally larger than the similarity score of the unmatched ones by a margin [10], [42]. As suggested in work [10], the loss regularization of most difficult negative samples often achieves better performance than the loss aggregation of all negative samples. Inspired by this finding, we focus on optimizing the hard negative samples that produce the highest loss and define the hard-negative triplet ranking loss $\mathcal{L}_{hntr}$ as

$$\mathcal{L}_{hntr} = \sum_{(\mathcal{V}, \mathcal{T})} \Big(\left[\alpha - S(\mathcal{V}, \mathcal{T}) + S\left(\mathcal{V}, \hat{\mathcal{T}}_-\right)\right]_+ + \left[\alpha - S(\mathcal{T}, \mathcal{V}) + S\left(\mathcal{T}, \hat{\mathcal{V}}_-\right)\right]_+ \Big) \tag{17}$$

where $[\cdot]_+ = \max(\cdot, 0)$ and α serves as a margin parameter. $\hat{\mathcal{T}}_-$ and $\hat{\mathcal{V}}_-$ are the hard negative samples, respectively, obtained by $\hat{\mathcal{T}}_- = \mathrm{argmax}_{\mathcal{T}_-} S(\mathcal{V}, \mathcal{T}_-)$ and $\hat{\mathcal{V}}_- = \mathrm{argmax}_{\mathcal{V}_-} S(\mathcal{V}_-, \mathcal{T})$, $\mathcal{T}_-$ and $\mathcal{V}_-$ are negative samples. In general, the hard-negative triplet ranking loss aims at enforcing the gap of similarity between the matched pairs and the unmatched paired converge to margin. However, it cannot ensure the similarity of matched pairs to be larger. To tackle this issue, as show in Fig. 2, we

Algorithm 1 Optimization Pseudocode for ReSG Framework

input: The image feature $\mathbf{H}^v$, text feature $\{\overrightarrow{\mathbf{H}}^t, \overleftarrow{\mathbf{H}}^t\}$;
1: Construct VSeG model $\mathbf{G}^v$ with initialized network parameter $\theta_{\mathbf{G}^v}$: $\mathbf{W}_l$, $\mathbf{b}_l$, $\mathbf{W}_1^v$, $\mathbf{b}_1^v$, $\mathbf{W}_2^v$, $\mathbf{b}_2^v$, $\mathbf{W}_c^v$, $\mathbf{b}_c^v$;
2: Construct TSeG model $\mathbf{G}^t$ with initialized network parameter $\theta_{\mathbf{G}^t}$: $\mathbf{W}_{e\_word}$,$\mathbf{W}_{e\_pos}$, $\overrightarrow{\mathbf{W}}_1^t$, $\overrightarrow{\mathbf{b}}_1^t$, $\overrightarrow{\mathbf{W}}_2^t$, $\overrightarrow{\mathbf{b}}_2^t$, $\overleftarrow{\mathbf{W}}_1^t$, $\overleftarrow{\mathbf{b}}_1^t$, $\overleftarrow{\mathbf{W}}_2^t$, $\overleftarrow{\mathbf{b}}_2^t$, $\overrightarrow{\mathbf{W}}_f^t$, $\overrightarrow{\mathbf{b}}^t$, $\overleftarrow{\mathbf{W}}_f^t$, $\overleftarrow{\mathbf{b}}^t$;
3: Initialize hyperparameters: k, α, δ, β, γ, λ_1, λ_2, η;
4: **repeat**
5: Calculate $\mathbf{A}^v$ via Eq. (3);
6: Calculate $\mathbf{A}^f$ via Eq. (10) and $\mathbf{A}^b$ via Eq. (11);
7: Update graph nodes $\mathcal{V}$ via Eq. (4) and $\mathcal{T}$ via Eq. (14);
8: Calculate loss $\mathcal{L}$ via Eq. (20);
9: Update $\theta_{\mathbf{G}^v}$ as: $\theta_{\mathbf{G}^v} = \theta_{\mathbf{G}^v} - \eta \frac{\partial \mathcal{L}}{\partial \theta_{\mathbf{G}^v}}$;
10: Update $\theta_{\mathbf{G}^t}$ as: $\theta_{\mathbf{G}^t} = \theta_{\mathbf{G}^t} - \eta \frac{\partial \mathcal{L}}{\partial \theta_{\mathbf{G}^t}}$;
11: **until** (convergency or reaching maximum iterations)
output: $\theta_{\mathbf{G}^v}$ and $\theta_{\mathbf{G}^t}$

design an auxiliary constraint $\mathcal{L}_{cen}$, called center hinge loss, to increase the absolute score of matched pairs while decreasing the absolute score of unmatched pairs

$$\mathcal{L}_{cen} = \sum_{(\mathcal{V},\mathcal{T})} \left(\sum_{\mathcal{V}_-} \left(\left[S(\mathcal{V}, \mathcal{T}_-) - \gamma \right]_+ \right) + \sum_{\mathcal{T}^-} \left(\left[S(\mathcal{T}, \mathcal{V}_-) - \gamma \right]_+ \right) + \left(\left[\beta - S(\mathcal{V}, \mathcal{T}) \right]_+ \right) \right) \tag{18}$$

where β and γ are hyperparameters. In addition, we impose another auxiliary constraint $\mathcal{L}_{pnm}$, called positive–negative margin loss, to increase the distance between the hard positive pairs and hard negative pairs in each modalities

$$\mathcal{L}_{pnm} = \sum_{(\mathcal{V},\mathcal{T})} \left(\left[\delta - S\left(\mathcal{V}, \hat{\mathcal{V}}_+\right) + S\left(\mathcal{V}, \hat{\mathcal{V}}_-\right) \right]_+ + \left[\delta - S\left(\mathcal{T}, \hat{\mathcal{T}}_+\right) + S\left(\mathcal{T}, \hat{\mathcal{T}}_-\right) \right]_+ \right) \tag{19}$$

where δ is a margin parameter. $\hat{\mathcal{T}}_+$ and $\hat{\mathcal{V}}_+$ are the hard positive samples, respectively, denoted as $\hat{\mathcal{T}}_+ = \text{argmin}_{\mathcal{T}} S(\mathcal{V}, \mathcal{T})$ and $\hat{\mathcal{V}}_+ = \text{argmin}_{\mathcal{V}} S(\mathcal{V}, \mathcal{T})$, $(\mathcal{V}, \mathcal{T})$. For efficient image–text matching, the regularization of different loss functions should be exploited in an integrated way, and the following objective function is utilized to learn the fine-grained correspondence:

$$\mathcal{L} = \mathcal{L}_{hntr} + \lambda_1 \mathcal{L}_{cen} + \lambda_2 \mathcal{L}_{pnm} \tag{20}$$

where λ_1 and λ_2 are balance parameters. Through the joint exploitation of (20), we utilize the Adam optimizer with 25 epochs for the optimization process, which can be iteratively solved until the convergence is reached. Let η be the learning rate in the Adam optimizer, the optimal parameters can be well obtained via Algorithm 1. Consequently, the semantic correspondence derived from image and text is semantically meaningful for benefiting various image–text matching tasks.

IV. Experiment

This section conducts a series of quantitative experiments to validate the efficiency of the proposed framework on fine-grained image–text matching task. The experiments and analysis will be detailed in the following sections.

A. Dataset and Evaluation Metric

Two public available multimodal datasets, that is, MSCOCO [43] and Flickr30K [44], are chosen in the experiments. MSCOCO contains 123 287 images, and each image is annotated with five captions. The widely used splitting scheme contains 113 287 images for training, 5000 images for validation, and 5000 images for testing [8]. Flickr30K contains 31 000 images collected from the Flickr website with five captions. Following the splitting scheme in [6] and [8], we select 1000 images for validation and 1000 images for testing and the rest for training. Meanwhile, the results for both 1k and 5k test sets are reported. In the case of 1k images, the results are averaged by performing a five-fold cross-validation on the 5k splitting test.

To quantitatively evaluate the matching performance, we report the score of Recall@K, which is the percentage of queries whose ground truth is ranked within top K instances, with higher score indicating the better performance [41]. Meanwhile, we also report "mR" score for overall evaluation, which averages all the recall values to assess the overall performance for both image-to-text matching and text-to-image matching tasks. In addition, mean average precision (mAP) [45], defined as the average AP of all queries, is also utilized for cross-modal matching evaluation. Accordingly, quantitative evaluation results of the image-to-text (I→T) matching and text-to-image (T→I) matching are reported.

B. Implementation Details

In the implementation, the proposed framework is implemented in the pytorch platform. For text representation learning, the word embedding size and the part-of-speech embedding size are, respectively, set at 300 and 15. The dimension of the word vector is fixed at 315, while the dimension of the joint embedding space D is set at 2048, the dimension d_a is set to 512, and the dimension d is set to 2048. For image representation learning, the pretrained visual features with 36 patches provided by SCAN [10] is selected for training, and each patch is characterized with 2048-dimension vector. Within the bidirectional graphs in text branch, k in (10) and (11) is set at 2. For the regularization parameters, we set the margin values as $\alpha = 0.2$ and $\delta = 0.7$, the hyperparameters as $\beta = 0.7$ and $\gamma = 0.3$, and the weights as $\lambda_1 = 0.3$ and $\lambda_2 = 0.1$. Similar to [32] and [46], we combine the results from two trained models by averaging their similarity scores and utilize the Adam optimizer with 20 epochs for the training process. Meanwhile, the initial learning rate is set at 0.0002, with decaying 10% every 5 epochs for both Flickr30k and MSCOCO datasets. In the experiments, we compare the proposed model with state-of-the-art competing methods, that is, global matching methods (WayNet [17], OEM [5], DSPE [42], DPC [27], GXN [28], and VSE++ [6]), local matching methods (CRAN [31], sm-LSTM [22], SCO [11], SCAN [10], CAMP [7], CASC [32], and MMCA [29]), and relationship matching methods (SGM [12], VSRN [46], and GSMN [13]). Alternatively, we also refer to the work [34]

TABLE I
IMAGE–TEXT MATCHING RESULTS ON THE FLIKER30K DATASET, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	Flickr30K Dataset						
	I→T Matching			T→I Matching			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
DSPE [42]	40.3	68.9	79.9	29.7	60.1	72.1	58.5
2WayNet [17]	49.8	67.5	-	36.0	55.6	-	52.2
GXN [28]	56.8	-	89.6	41.5	-	80.1	67.0
VSE++ [6]	52.9	80.5	87.2	39.6	70.1	79.5	68.3
DPC [27]	55.6	81.9	89.5	39.1	69.2	80.9	69.4
sm-LSTM [22]	42.5	71.9	81.5	30.2	60.4	72.3	58.8
CRAN [31]	38.1	71.1	82.1	38.1	70.8	82.8	63.8
DAN [9]	55	81.8	89.0	39.4	69.2	79.1	68.9
SCO [11]	55.5	82.0	89.3	41.4	70.5	80.1	69.8
SCAN [10]	67.4	90.3	95.8	48.6	77.7	85.2	77.5
CAMP [7]	68.1	89.7	95.2	51.5	77.1	85.3	77.8
CASC [32]	68.5	90.6	95.9	50.2	78.3	86.3	78.3
MMCA [29]	74.2	92.8	96.4	54.8	81.4	87.8	81.2
SGM[12]	71.8	91.7	95.5	53.5	79.6	86.5	79.8
VSRN [46]	71.3	90.6	96.0	54.7	81.8	88.2	80.5
GSMN (sparce) [13]	71.4	92.0	96.1	53.9	79.7	87.1	80.0
GSMN (dense) [13]	72.6	93.5	96.8	53.7	80.0	87.0	80.6
GSMN (sparce+dense) [13]	76.4	**94.3**	97.3	57.4	82.3	89.0	82.8
ReSG	74.8	94.0	97.3	57.2	82.4	**89.1**	82.5
ReSG*	**77.2**	94.2	**98.2**	**58.0**	**83.1**	88.7	**83.2**

and select the pretrained 300-dim GloVe vector trained on the Wikipedia dataset to initialize the text feature vector. The results obtained by this way are marked as "ReSG*."

C. Results of Image–Text Matching Performance

The image–text matching results tested on different datasets are shown in Tables I and II, 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 these global matching methods. Specifically, WayNet utilizes generative models to extract the global features of images and texts, which generally ignores the salient image structure embedded in real-world data and often degrades its performance in practice. In addition, the R@1 scores I→T task obtained by the GXN [28] method is only equal to 56.8 and 68.5, respectively, tested on Flickr30K and MSCOCO 1K datasets, which are relatively poor for real applications. Comparatively speaking, the local matching and relationship matching methods are able to deliver better image–text matching performance. For instance, SCAN [10] and MMCA [29] both employ the cross-model attention to attend salient regions and key words and capture the fine-grained interplay between vision and language, which generally performs better than the global matching methods. For instance, the R@5 scores of I→T task obtained by the MMCA method are reached up to 92.8 and 95.6, respectively, tested on Flickr30K and MSCOCO 1K datasets. This indicates that the MMCA method is capable of returning much more similar samples in the retrieval results, which plays an important role for a practical retrieval system.

Further, the performances delivered by SGM, VSRN, and GSMN methods are generally better than that obtained by global matching methods (e.g., GXN [28] and VSE++ [6]) and local matching methods (e.g., SCAN [10] and CAMP [7]). That is, the high-level semantic relationships can provide valuable information for fine-grained image–text matching. It is noted that SGM, VSRN, and GSMN methods also explore the higher-order concepts and their semantic relationship. Nevertheless, SGM and GSMN methods only consider one directional relationship for textual data, while the VSRN approach only reasons the relationships of image patches. In contrast to this, the proposed ReSG framework improves the textual graph representation by extracting bidirectional textual semantic relationship, while considering more discriminative loss function to learn the fine-grained semantic correspondence. As shown in Table I, it can be observed that the proposed ReSG framework has yielded comparable and even better performances than that obtained by other baselines. Specifically, the proposed ReSG* framework outperforms the state-of-the-art baselines by achieving the best R@1 scores on different datasets. For instance, the proposed ReSG* framework outperforms the global matching methods and local matching methods by a large margin and also gains the R@1 improvements of 5.9% at I→T matching task and 4.1% at T→I matching task in comparison with the VSRN method. This indicates that the proposed framework is capable of indexing much more similar samples in the cross-modal matching results. Although the R@5 score of I→T task obtained by the proposed framework and tested on the Fliker30K dataset is slightly lower than that obtained by the GSMN [13] method, our proposed framework always delivers the best mean recall scores in all retrieval tasks.

Besides, we further utilize mAP@K values to measure the cross-modal retrieval performances. For mAP@K metric, the larger value generally indicates the better cross-modal matching performance. Since the mAP results obtained by most competing works are not reported in their original papers and their source codes are not released currently, we compare the proposed framework with four state-of-the-art baselines, that is, VSE++ [6], SCAN [10], CAMP [7], and VSRN [46]. As shown in Table III, it can be observed that the proposed approach has yielded the better image–text retrieval performance than that obtained by the competing baselines. For instance, the mAP@10 scores of I→T task obtained by the CAMP [7] and VSRN [46] methods, respectively, reach to 65.2% and 70.0%, when tested on the Flickr30K dataset. In contrast, the mAP@10 score of I→T task obtained by the proposed approach reaches up to 73.1% when evaluated on the Flickr30K dataset. That is, the proposed framework performs well in fine-grained cross-modal retrieval tasks. The main superiorities contributed to these very competitive performances are two-fold.

1) The proposed relationship-enhanced graph model is beneficial to capture fine-grained correspondence between image and text data. Accordingly, the derived relation correspondence is able to guide the fine-grained object correspondence learning, while the fine-grained object correspondence simultaneously forces the network to learn relation correspondence explicitly.
2) The designed loss function is able to well learn the fine-grained object correspondence and relation correspondence. Consequently, the derived relationship-enhanced

TABLE II
QUANTITATIVE IMAGE–TEXT MATCHING RESULTS ON THE MSCOCO TEST SET, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	MSCOCO 1K							MSCOCO 5K						
	I→T Matching			T→I Matching			mR	I→T Matching			T→I Matching			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	
2WayNet [17]	39.7	63.3	-	55.8	75.2	-	58.5	-	-	-	-	-	-	-
OEM [5]	46.7	-	88.9	37.9		85.9	64.8	23.3	-	84.7	31.7	-	74.6	53.6
DSPE [42]	50.1	79.7	89.2	39.6	75.2	86.9	70.1	-	-	-	-	-	-	-
DPC [27]	65.6	89.8	95.5	47.1	79.9	90.0	78.0	41.2	70.5	81.1	25.3	53.4	66.4	56.3
GXN [28]	68.5	-	97.9	56.6	-	94.5	79.4	42.0	-	84.7	31.7	-	74.6	58.25
VSE++ [6]	64.6	90.0	95.7	52.0	84.3	92.0	79.8	41.3	71.1	81.2	30.3	59.4	72.4	59.3
CRAN [31]	-	-	-	-	-	-	-	23.0	52.0	66.0	21.1	48.9	64.5	46.0
sm-LSTM [22]	53.2	83.1	91.5	40.7	75.8	87.4	72.0	-	-	-	-	-	-	-
SCO [11]	69.9	92.9	97.5	56.7	87.5	94.8	83.2	42.8	72.3	83.0	33.1	62.9	75.5	61.6
SCAN [10]	72.7	94.8	98.4	58.8	88.4	94.8	84.7	50.4	82.2	90.0	38.6	69.3	80.4	68.5
CAMP [7]	72.3	94.8	98.3	58.5	87.9	95.0	84.5	50.1	82.1	89.7	39.0	68.9	80.2	68.3
CASC [32]	72.3	96.0	**99.0**	58.9	89.8	**96.0**	85.3	47.2	78.3	87.4	34.7	64.8	76.8	64.9
MMCA [29]	74.8	95.6	97.7	61.6	89.8	95.2	85.8	54.0	82.5	90.7	38.7	69.7	80.8	69.4
SGM [12]	73.4	93.8	97.8	57.5	87.3	94.3	84.0	50.0	79.3	87.9	35.3	64.9	76.5	65.7
VSRN [46]	76.2	94.8	98.2	62.8	89.7	95.1	86.1	53.0	81.1	89.4	40.5	70.6	81.1	69.2
GSMN(sparce) [13]	76.1	95.6	98.3	60.4	88.7	95.0	85.7	-	-	-	-	-	-	-
GSMN (dense) [13]	74.7	95.3	98.2	60.3	88.5	94.6	85.3	-	-	-	-	-	-	-
GSMN (sparce+dense) [13]	78.4	96.4	98.6	63.3	90.1	95.7	87.1	-	-	-	-	-	-	-
ReSG	78.1	96.2	98.0	64.1	**90.5**	**96.0**	87.1	55.1	82.5	90.3	41.8	**72.7**	82.0	70.7
ReSG*	**79.3**	**96.7**	98.3	**64.5**	90.0	95.8	**87.2**	**55.8**	**83.0**	**91.0**	**42.0**	72.4	**82.1**	**71.1**

TABLE III
QUANTITATIVE COMPARISONS OF IMAGE–TEXT MATCHING PERFORMANCE (MAP@K), AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. FOR SIMPLICITY, "*" IS THE ABBREVIATED FORM OF "MAP" IN THE TABLE

Method	Flickr30K 1K test								MSCOCO 1K test							
	I→T Matching				T→I Matching				I→T Matching				T→I Matching			
	mAP@1	*@10	*@100	*@All	*@1	*@10	*@100	*@All	mAP@1	*@10	*@100	*@All	mAP@1	*@10	*@100	*@All
VSE++ [6]	52.9	50.9	42.8	39.8	39.6	51.2	51.9	51.9	64.6	60.1	54.3	52.0	52.0	60.5	61.1	61.1
SCAN [10]	67.4	64.0	53.1	49.6	48.6	57.0	57.7	57.7	72.7	71.9	62.9	62.2	58.8	70.4	70.7	70.7
CAMP [7]	67.1	65.2	54.4	50.1	50.3	59.6	60.5	60.5	71.2	70.8	64.3	64.3	58.1	69.7	69.8	69.8
VSRN [46]	71.4	70.0	60.0	57.9	54.0	65.4	66.0	66.0	76.1	76.0	68.7	67.6	62.5	75.8	76.0	76.0
ReSG	**74.8**	**73.1**	**62.4**	**61.0**	**57.2**	**67.5**	**69.2**	**69.2**	**78.1**	**76.3**	**70.7**	**69.4**	**64.1**	**76.8**	**77.4**	**77.4**

graph representations are more semantically meaningful for the efficient image–text matching and retrieval tasks.

D. Ablation Studies

Within the proposed ReSG framework, two tailored graph models (i.e., VReG and TReG) and discriminative loss functions are carefully considered for the efficient image–text matching. Next, we further evaluate the effectiveness of each learning module and validate the performance of different learning combinations.

1) *Base:* We remove the bidirectional relationship extraction in the text branch and semantic relationship extraction in the image branch, and ignore the auxiliary loss and part-of-speech information.
2) *VReG:* Extension of the "base" model by adding the semantic relationship extraction in the image branch.
3) *TReG:* Extension of the base model by adding the bidirectional relationship extractor in the textual branch.
4) *Pos:* Extension of the base model by adding the part-of-speech information into words.
5) $\mathcal{L}_{pnm}$*:* Extension of the base model by adding the positive–negative margin loss.
6) $\mathcal{L}_{cen}$*:* Extension of the base model by adding the center hinge loss.

The detailed combinations are shown in Table IV, in which $\surd$ means the embedding of such module, *bg* and *fg*, respectively, denote the textual graph with backward and forward relationship embeddings. It can be clearly observed that the embedding of part-of-speech information, relationship-enhanced graph, and auxiliary loss constraints has significantly improved the image–text matching performances. From tasks (4, 5, 6) or (9, 10, 11), it can be observed that if we only leverage one auxiliary loss, the image–text matching results may not be improved or even dropped, while the proposed model performs much better if both of the positive–negative margin loss and center hinge loss are embedded. From tasks (6, 11, 12) or (15, 16, 17), it can be found that the embedding of relationship-enhanced image features also yields the significant improvements, while the embedding of the bidirectional textual relationship can improve the matching performance to some degree. That is, the embedding of ReSG and the designed loss functions are able to boost the image–text matching performance.

Besides, we show the training times obtained by various module combinations and different competing baselines, that is, SCAN, VSRN, and GSMN. The model is trained on the GPU NVIDIA RTX 2080Ti, the batch size within the SCAN and VSRN methods is set at 128, while the batch sizes within

TABLE IV
ABLATION STUDIES TESTED ON THE FLICKR30K DATASET (1K TEST), AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Ablation Study on Flickr30K												
Index	Components						I→T Matching			T→I Matching		
	base	**VReG**	**TReG**	**pos**	$\mathcal{L}_{pnm}$	$\mathcal{L}_{cen}$	R@1	R@5	R@10	R@1	R@5	R@10
1	√						67.4	90.1	93.9	52.0	78.8	86.0
2	√	√					71.7	90.3	95.4	53.4	80.5	87.4
3	√	√		√			71.6	91.2	94.9	53.9	80.2	88.0
4	√	√		√	√		72.0	91.1	95.1	54.3	80.7	87.7
5	√	√		√		√	71.9	91.3	94.8	54.4	80.4	88.0
6	√	√		√	√	√	72.4	91.2	95.8	54.6	81.1	87.9
7	√		√				68.7	90.6	94.2	53.1	79.3	87.2
8	√		√	√			68.9	90.4	94.4	53.8	79.2	87.5
9	√		√	√	√		69.2	90.2	94.3	53.5	81.1	87.9
10	√		√	√		√	70.2	91.0	94.7	54.2	80.8	87.4
11	√		√	√	√	√	70.4	90.7	94.7	54.5	80.4	87.0
12	√	√	√	√			73.2	92.4	96.1	55.4	81.8	88.1
13	√	√	√	√	√		73.0	92.4	96.0	54.9	81.0	87.1
14	√	√	√	√		√	74.0	93.0	96.4	56.3	82.1	88.6
15	√	√	bg	√	√	√	73.1	92.8	96.2	55.4	81.3	88.2
16	√	√	fg	√	√	√	73.8	93.3	96.8	56.0	81.5	88.6
17	√	√	√	√	√	√	**74.8**	**94.0**	**97.3**	**57.2**	**82.4**	**89.1**

TABLE V
TRAINING TIMES OBTAINED BY DIFFERENT APPROACHES AND TESTED ON THE FLICKR30K DATASET

Methods	Traing Time (hours)	mR (%)
SCAN [10]	4.7	77.5
VSRN [46]	6.6	79.4
GSMN (sparce) [13]	12.1	80.0
GSMN (dense) [13]	13.5	80.6
ReSG (base model)	2.8	78.0
ReSG (base&VReG)	5.0	79.8
ReSG (base&VReG&TReG)	6.6	81.0
ReSG (base&VReG&TReG&pos)	7.0	81.2
ReSG (full model)	7.5	82.5

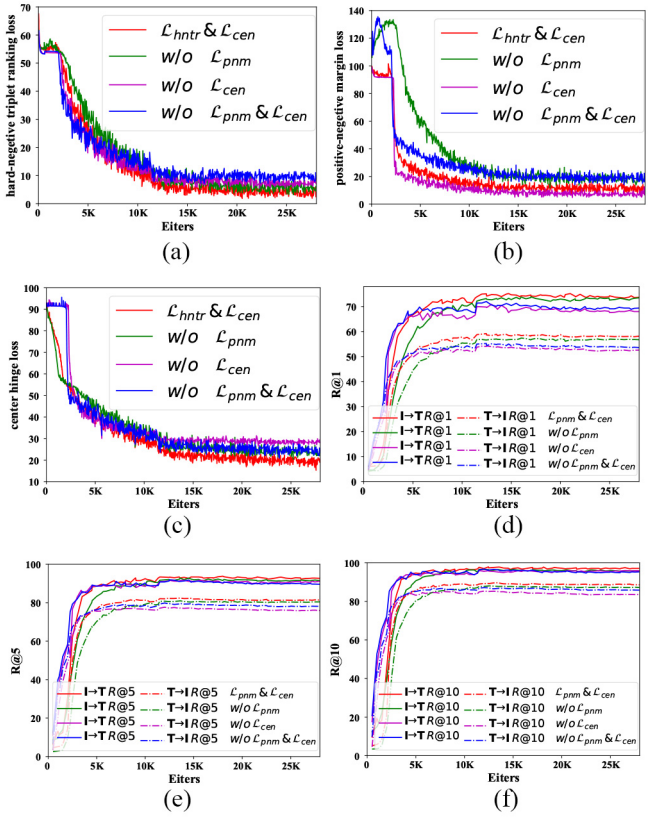

Fig. 4. Illustration of variations under different loss functions. (a) $\mathcal{L}_{hntr}$ loss. (b) $\mathcal{L}_{pnm}$ loss. (c) $\mathcal{L}_{cen}$ loss. (d) R@1. (e) R@5. (f) R@10.

the GSMN and the proposed ReSG approach are fixed to be 64 and 100, respectively. As illustrated in Table V, the proposed ReSG method achieves a good balance between the time cost and image–text matching performance. Remarkably, the proposed ReSG approach is running much faster than the GSMN method. The main reason lies in that GSMN often involves large iterations to convergence. Since the proposed ReSG approach considers more modules and loss functions to discriminatively learn the relationship-enhanced graph representations, the execution time of training time could be much higher than that obtained by the SCAN and VSRN methods. Fortunately, the proposed ReSG method does not significantly increase the training time to a large extent, while achieving the best image–text retrieval performances. Therefore, the proposed ReSG approach is suitable for processing large-scale image–text retrieval tasks from a practical viewpoint.

Further, we draw the loss curves to verify the validity of the designed auxiliary loss functions. To be specific, we monitor the variations of each loss function by adding or dropping the positive–negative margin loss or center hinge loss from the framework. As shown in Fig. 4, the blue curve in subfigure (a) shows the change of the positive–negative margin loss $\mathcal{L}_{pnm}$ under $\lambda_2 = 0$ (i.e., the positive–negative margin loss is not embedded into the model), while the red curve shows the change of the positive–negative margin loss under $\lambda_2 = 0.1$ (i.e., the positive–negative margin loss is embedded into the model). In subfigure (b), the green curve shows the change of the center hinge loss $\mathcal{L}_{cen}$ under $\lambda_1 = 0$ (i.e., $\mathcal{L}_{cen}$ is not embedded into the model), while the red curve shows the change of the center hinge loss $\mathcal{L}_{cen}$ under $\lambda_1 = 0.2$ (i.e., $\mathcal{L}_{cen}$ is embedded into the model). After the loss converges, it can be observed the positive–negative margin loss

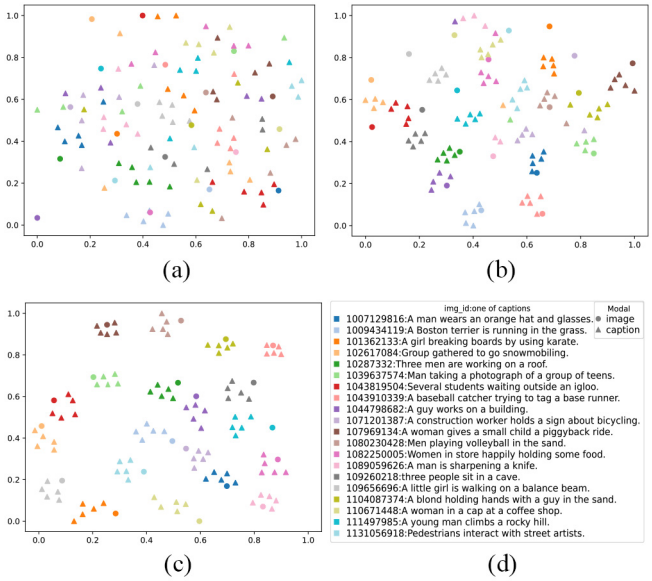

Fig. 5. Illustration of feature distance between 20 image examples and 100 captions (each image is annotated with five captions). (a) Before training. (b) Base model. (c) Full model. (d) Cluster examples.

described by the red curve is converged at a lower rate than the loss described by the blue curve, and the similar results can be also found in the center hinge loss. This indicates that the distances between the semantically irrelevant samples are enlarged within the same modality, while the semantically relevant samples are becoming closer. Meanwhile, it can be also observed from R@1 results that the proposed model embedded with positive–negative margin loss and the center hinge loss often performs better that the model without corresponding embeddings.

E. Visualization Analysis

To further demonstrate the interpretability of the proposed model, we utilize the t-SNE algorithm to visualize the embeddings of 20 images and 100 corresponding captions on the Flickr30K dataset. As shown in Fig. 5, it can be found that the base model is able to cluster some similar images and captions together, but which may have some overlaps among different semantic representations. For instance, the image example with ID "1089059626" and its relevant captions do not form into an intensive cluster. In contrast to this, the proposed ReSG framework can well put the feature embedding of two similar modalities close together while pulling those of different modalities away, and the derived feature embeddings are discriminative for various image–text matching tasks.

Further, we show the representative examples in Flickr30K 1K test data to visualize the learned components. More specifically, we calculate the sum of weights of all the edges connected to the nodes, and then utilize different colors to visualize them on the detected regions. Representative examples are shown in Fig. 6(e), in which the sum of the weights of the edges connected to the current region is ranked in the top 15, and the warmer red indicates that the area aggregates more relationship information from other objects. It can be observed that the proposed graph model can better aggregate most informative relationships into the representation of salient image objects, such as *racket*, *woman*, and *people behind her* in the first row, and *ball* and *soccer players* in the second row. Further, we visualize the weighted adjacency matrices of the forward and backward textual graph encoders, respectively, shown in column (b) and column (c). It can be clearly observed that the relationships between each pair of words are quantitatively weighted in the forward topology graph and the backward topology graph. On the one hand, most of the words in the sentence, such as *the*, *with*, and *as*, are less informative, and their aggregated weights are very small. This indicates that the proposed textual model can well filter the redundant information in the sentence. On the other hand, most of the attribute words, instance words, and verbs are semantically correlated with each other, for example, the semantic relationships of most informative words are retained in the backward graph. Therefore, the backward graph is valuable to provide significant relationship information for discriminative representation.

Besides, we further show some representative examples of the proposed model in fine-grained image–text matching. As shown in Fig. 7, the upper parts show the I→T matching results specified by image query, while the lower parts display the T→I matching results specified by text query. For I→T matching, two representative groups are presented, and each group contains three very similar image queries associated with the ranked matching results. For T→I matching, we also show two representative groups, and each group contains two similar text queries associated with the ranked matching results. From the retrieval results, it can be clearly observed that the proposed model is able to distinguish the similar queries well and have successfully indexed the most semantically matched counterparts. For instance, on the one hand, the text instances containing *barking* are successfully retrieved in the second row, while the text examples containing *holding* and *fish* are successfully indexed in the third row. On the other hand, the images with semantic concepts *swinging* and *holding*, which also appeared in the text queries, are also indexed successfully. That is, the proposed framework is capable of capturing the fine-grained relationships between the images and texts, leading to the outstanding matching performances.

F. Parameters Analysis

Within the proposed learning framework, several parameters are involved, that is, λ_1, λ_2, the number of VReG layers, the number of TReG layers, and the value of k in TReG. Next, we select the Flickr30K dataset for evaluation and conduct extensive experiments with different parameter values to investigate the effect of these hyperparameters. As shown in Fig. 8, we first vary the values of λ_1 and λ_2 in (20) with different values (i.e., 0, 0.1, 0.3, 0.5, and 0.7), and record the R@1 values in both I→T and T→I matching scenarios. It can be found that the results perform well when λ_1 is selected within the range of [0.1, 0.5] and λ_2 is chosen within the range of [0.1, 0.3]. In the experiments, the settings of $\lambda_1 = 0.3$ and $\lambda_2 = 0.1$ often deliver the competitive performances.

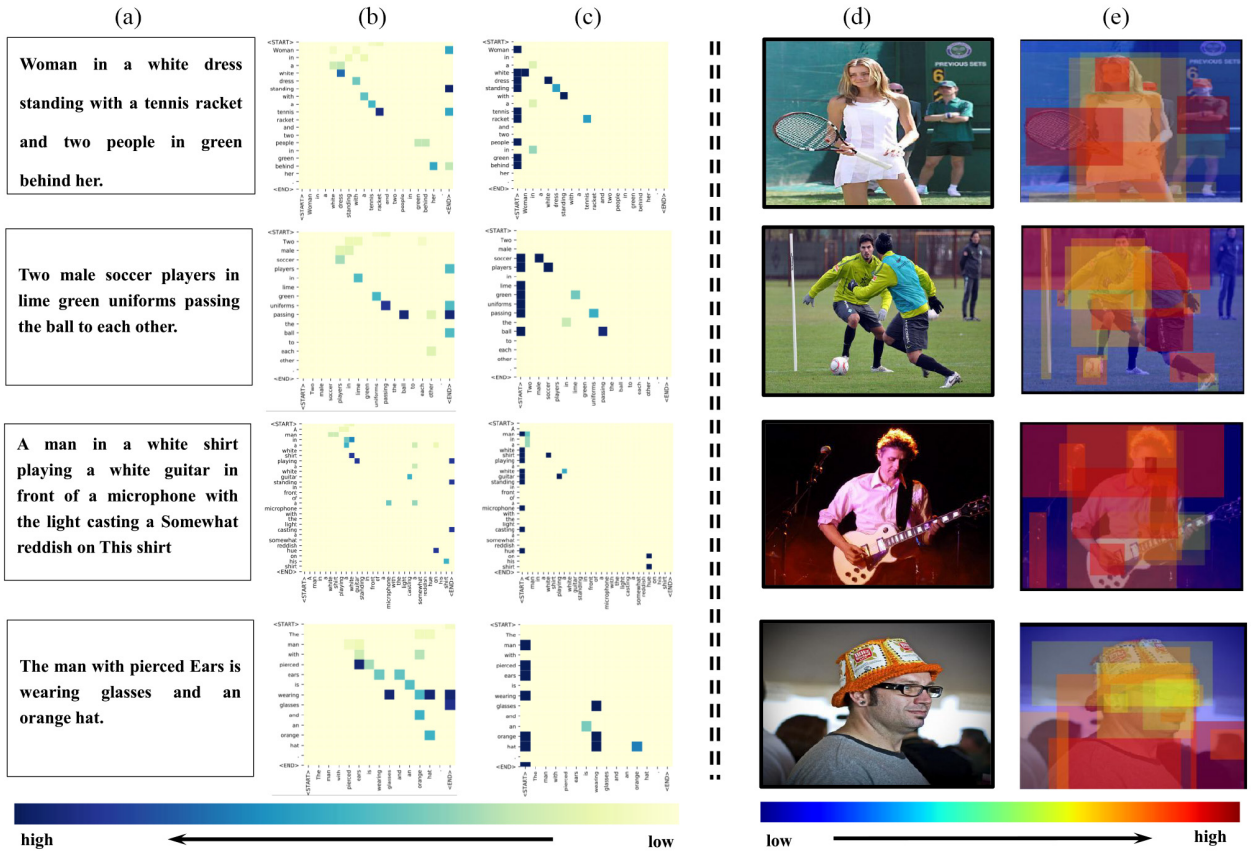

Fig. 6. Visualization of semantic relationships, and each row is a matched image–text instance. (a) Raw text data. (b) Forward weights in textual graph. (c) Backward weights in textual graph. (d) Raw image data. (e) Visualization of the summed graph edges.

Fig. 7. Representative examples of cross-modal matching results between images and texts. For ease of reference, some objects and attributes are marked as blue in textual sentences, while the verbs are marked as red.

Moreover, we further explore the effect of feature extraction layers within the proposed VReG and TReG models, and assess the parameter k that influences the nearest neighbor number in (10) and (11). Representative results tested on different values are shown in Fig. 9, it can be observed that the settings of the TReG layer, VReG layer, and k nearest

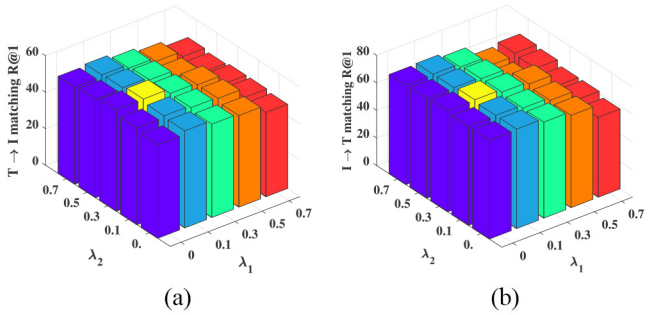

Fig. 8. Evaluation of different λ_1 and λ_2 values on the Flickr30k dataset, and the best results are marked with yellow color. (a) T→I matching results. (b) I→T matching results.

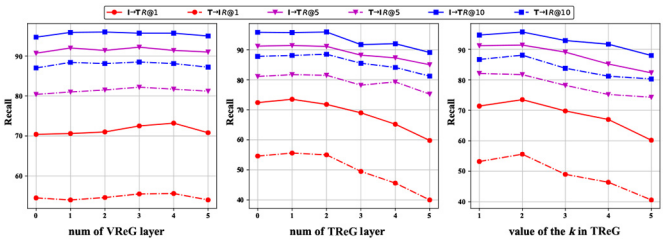

Fig. 9. Evaluation of VReG and TReG layer numbers, and nearest neighbor number k on image retrieval.

neighbors, respectively, within the range of [1, 4], [1, 3], and [1, 3], only induce a minor fluctuation to the cross-modal retrieval performance. Specifically, the R@1 and R@5 results become better until the numbers of VReG layers, TReG layers, and the value of k are, respectively, equal to 4, 2, and 2. Therefore, these parameters are generally insensitive to the image–text retrieval performances within a wide range of values.

V. Conclusion

In this article, we have proposed an efficient relationship-enhanced graph model to achieve fine-grained image–text matching. Within the proposed framework, two tailored graph encoders, VReG and TReG, are, respectively, exploited to encode the high-level semantic concepts of corresponding instances and their contextual semantic relationships. Specifically, the TReG encoder embeds the part-of-speech information into node representation and considers forward–backward topologies to discriminatively characterize the relationship-enhanced textual features. Meanwhile, the representations of each node on these graph models are optimized by aggregating semantically contextual information, while the hard-negative triplet ranking loss, center hinge loss, and positive–negative margin loss are seamlessly integrated to jointly learn the fine-grained correspondence between the designed image and text graph representations. Accordingly, the derived relationship-enhanced features aggregated in these graph models can be well utilized for image–text matching in a more interpretable and plausible way. Extensive experiments evaluated on various kinds of image–text matching tasks have shown its outstanding performance.

Along the line of the present work, several open problems also deserve our further research. For example, the current graph model often attempts to enhance the representations of each node by aggregating useful concept information from other nodes within each modality individually, which may loss some local-graph correspondence across heterogeneous modalities. Theoretically, it is also beneficial to pay more attention on some informative cross-node relationships between some salient regions and key words in a node-level fashion. Besides, the salient object detection methods would also have an influence on the image–text matching results, and more robust object detection methods deserve further investigation. We shall leave these studies in our future works.

References

[1] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2623–2631.

[2] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[3] Q. Wu, C. Shen, P. Wang, A. Dick, and A. V. D. Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[4] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *Int. J. Comput. Vis.*, vol. 125, nos. 1–3, pp. 110–135, 2017.

[5] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–10.

[6] 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. 1–13.

[7] Z. Wang *et al.*, "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5764–5773.

[8] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[9] H. Nam, J. W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2156–2164.

[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] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.

[12] 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.*, 2020, pp. 1497–1506.

[13] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10921–10930.

[14] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[15] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2160–2167.

[16] 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.

[17] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4601–4611.

[18] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 964–981, Mar. 2021.

[19] J. Zhang, Y. Peng, and M. Yuan, "SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 489–502, Feb. 2020.

[20] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[21] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 7–16.
[22] 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.*, 2017, pp. 7254–7262.
[23] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7181–7189.
[24] J. Wehrmann and R. C. Barros, "Bidirectional retrieval made simple," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7718–7726.
[25] 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.
[26] Y. Peng and J. Qi, "Reinforced cross-media correlation learning by context-aware bidirectional translation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1718–1731, Jun. 2020.
[27] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 2, pp. 1–23, 2020.
[28] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7181–7189.
[29] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 941–950.
[30] Y. Wu, S. Wang, and Q. Huang, "Learning semantic structure-preserved embeddings for cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 825–833.
[31] J. Qi, Y. Peng, and Y. Yuan, "Cross-media multi-level alignment with relation attention network," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 892–898.
[32] 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.
[33] Y. Peng, J. Qi, and Y. Zhuo, "MAVA: Multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism," *IEEE Trans. Image Process.*, vol. 29, pp. 2728–2741, 2020.
[34] H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, "Consensus-aware visual-semantic embedding for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 18–34.
[35] J. A. Pine, G. Csurka, and S. Clinchant, "Unsupervised visual and textual information fusion in CBMIR using graph-based methods," *ACM Trans. Inf. Syst.*, vol. 33, no. 2, pp. 1–31, 2015.
[36] Y. Yuan, Z. Xiong, and Q. Wang, "ACM: Adaptive cross-modal graph convolutional neural networks for RGB-D scene recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9176–9184.
[37] Y. Liu, B. Wan, X. Zhu, and X. He, "Learning cross-modal context graph for visual grounding," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 645–652.
[38] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.
[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[40] T. J. Fu, P. H. Li, and W. Y. Ma, "GraphRel: Modeling text as relational graphs for joint entity and relation extraction," in *Proc. Assoc. Comput. Linguist.*, 2020, pp. 1409–1418.
[41] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 707–723.
[42] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5005–5013.
[43] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[44] 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. Linguist.*, vol. 2, pp. 67–78, Dec. 2014.
[45] 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.
[46] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4654–4662.

Xin Liu (Senior Member, IEEE) received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, 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 Zhejiang Lab, Nanjing, China. He has authored over 50 papers in well-known international journals and conferences. His current research interests include multimedia data analysis, pattern recognition, and machine learning.

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

He is currently with the Department of Computer Science, Huaqiao University, where he is also a Research Fellow with the Xiamen Key Laboratory of Computer Vision and Pattern Recognition and Fujian Key Laboratory of Big Data Intelligence and Security. His current research interests include multimedia data analysis, pattern recognition, 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. His current research interests include machine learning, pattern recognition, and visual computing.

Prof. Cheung is the Founding Chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He serves as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, *Pattern Recognition*, *Knowledge and Information Systems*, and *Neurocomputing*. He is an IET Fellow, an RSA Fellow, and a BCS Fellow.

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 of Science and Technology of China, Chengdu, China. His current research interests include information retrieval, pattern recognition, and computer vision.

Nannan Wang (Member, IEEE) received the B.Sc. degree in information and computation science from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2009, and the Ph.D. degree in information and telecommunications engineering from Xidian University, Xi'an, in 2015.

He is currently a Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published over 150 articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, CVPR, and ICCV. His current research interests include computer vision, pattern recognition, and machine learning.