Learning Aligned Image-Text Representations Using Graph Attentive Relational Network

Ya Jing, Wei Wang, Liang Wang, *Fellow, IEEE*, and Tieniu Tan, *Fellow, IEEE*

***Abstract*— Image-text matching aims to measure the similarities between images and textual descriptions, which has made great progress recently. The key to this cross-modal matching task is to build the latent semantic alignment between visual objects and words. Due to the widespread variations of sentence structures, it is very difficult to learn the latent semantic alignment using only global cross-modal features. Many previous methods attempt to learn the aligned image-text representations by the attention mechanism but generally ignore the relationships within textual descriptions which determine whether the words belong to the same visual object. In this paper, we propose a graph attentive relational network (GARN) to learn the aligned image-text representations by modeling the relationships between noun phrases in a text for the identity-aware image-text matching. In the GARN, we first decompose images and texts into regions and noun phrases, respectively. Then a skip graph neural network (skip-GNN) is proposed to learn effective textual representations which are a mixture of textual features and relational features. Finally, a graph attention network is further proposed to obtain the probabilities that the noun phrases belong to the image regions by modeling the relationships between noun phrases. We perform extensive experiments on the CUHK Person Description dataset (CUHK-PEDES), Caltech-UCSD Birds dataset (CUB), Oxford-102 Flowers dataset and Flickr30K dataset to verify the effectiveness of each component in our model. Experimental results show that our approach achieves the state-of-the-art results on these four benchmark datasets.**

***Index Terms*— Image-text matching, cross-modal retrieval, person search, graph neural network.**

I. Introduction

JOINTLY learning vision and language is a significant task in the computer vision and pattern recognition community, which has drawn great attention in recent years. There are various research tasks in this field, e.g., image-text retrieval [1]–[3], visual question answering [4], [5], and image captioning [6], [7]. Great progress has been made with the development of deep learning. Despite these advances, cross-modal matching remains to be solved due to the semantic gap between vision and language. In this paper, we study the task of identity-aware image-text matching which aims to search images of the same identity as text queries and retrieve texts describing the same identity as image queries.

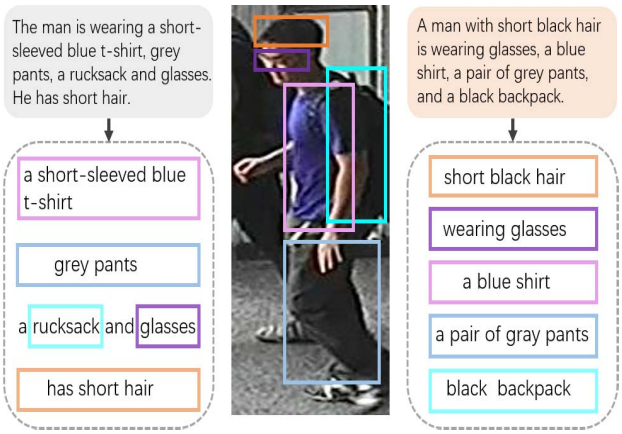

Fig. 1. Challenges in identity-aware image-text matching. (1) The misalignment between textual descriptions that describe the same image due to the sentence structure variation. (2) The latent semantic alignment between image regions and noun phrases needs to be reasoned.

However, there are several challenges for this task. First, complex relations between language descriptions and image appearances are highly non-linear, e.g., the corresponding relations between noun phrases and image regions. Second, people generally describe the same image with different orders of descriptions due to their different concerns. As seen in Fig. 1, both sentences describe the middle image, but they are not aligned well. The left text describes the t-shirt first while the right one describes the hair first. Due to the recurrent encoding manner in texts, different sentence structures will result in different textual features though they have the same semantic meaning. In a word, directly using the unaligned features which do not explore the sematic alignment between image and text for matching is not suitable. Therefore, the challenge of this task lies in learning aligned cross-modal features. Motivated by the similar observations, some prior methods propose to use the attention mechanism to match image regions with text words. Li *et al.* [2] propose a co-attention approach which includes spatial attention and latent semantic attention to learn the aligned features. Lee *et al.* [1] utilize

Manuscript received November 24, 2019; revised October 12, 2020 and December 5, 2020; accepted December 14, 2020. Date of publication January 8, 2021; date of current version January 18, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001000, in part by the National Natural Science Foundation of China under Grant 61976214 and Grant 61721004; and in part by the Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) under Grant 2019JZZY010119. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Giulia Boato. *(Corresponding author: Wei Wang.)*

Ya Jing and Wei Wang are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100044, China (e-mail: ya.jing@cripac.ia.ac.cn; wangwei@nlpr.ia.ac.cn).

Liang Wang and Tieniu Tan are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100044, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China (e-mail: wangliang@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2020.3048627

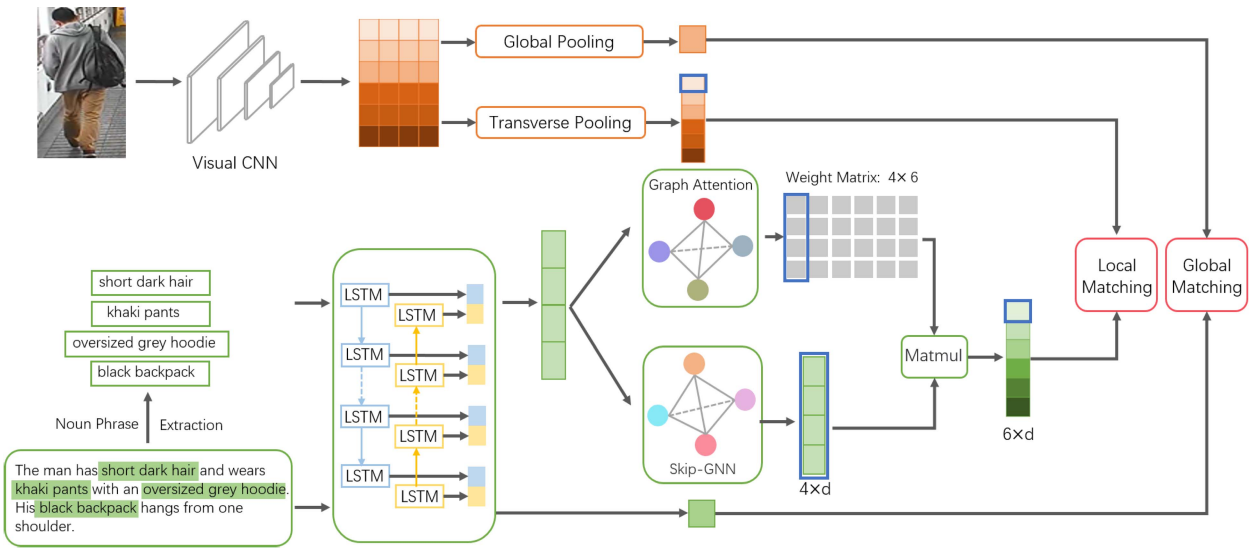

Fig. 2. The framework of our proposed graph attentive relational network (GARN). We utilize a skip graph neural network (skip-GNN) to learn effective textual representations which have relational features in addition to textual features. To further learn the latent semantic alignment between image regions and noun phrases, we propose a graph attention network to obtain the probabilities that the noun phrases belong to the image regions. The global and local matchings are utilized to supervise the learning of visual and textual representations. Both ranking loss and identification loss are employed to train our model, which aims to minimize the intra-identity distance and maximize the inter-identity distance simultaneously.

a stacked cross attention between detected image regions and words in sentences to infer the latent semantic alignment. But these attention methods regard different words in sentences as individuals and ignore the relationships between words which determine whether the words belong to the same visual object.

To solve the problems above, here we propose a graph attentive relational network (GARN) to learn the aligned image-text representations by modeling the relationships between local textual features. The framework of our model is shown in Fig. 2. We first utilize a visual convolutional neural network (CNN) [8] to extract visual feature maps. Then we obtain horizontal representations by horizontal pooling. For the textual input, we first extract noun phrases and then a bi-directional long short-term memory (LSTM) [9] network is employed to learn textual features. With the features of noun phrases, a skip graph neural network is proposed, where nodes in the graph represent the noun phrases in sentence and edges represent the relationships between the nodes. This skip graph neural network can learn a more effective textual representation by combining the textual features with the relational features. To learn the aligned image-text representations, we propose a graph attention network to learn the corresponding relationships between image regions and noun phrases. This attention network learns the probabilities that the noun phrases belong to the image regions by modeling the relationships between noun phrases. When training the model, we perform not only the global matching but also the local matching to learn more discriminative representations. In addition, both pair-wise ranking loss and identification loss are used to jointly minimize the intra-identity distance and maximize the inter-identity distance. To demonstrate the effectiveness of the proposed model, we perform experiments on four identity-aware cross-modal matching datasets: CUHK Person Description (CUHK-PEDES) [10], Caltech-UCSD Birds (CUB) [3], Oxford-102 Flowers [3] and Flickr30K [11], and achieve the state-of-the-art results.

The main contributions of our work are four-fold:

- We propose a novel graph attentive relational network (GARN) to learn the aligned image-text representations.
- The novel skip graph neural network aims to learn effective textual representations by integrating textual features with relational features.
- We model the latent visual-semantic alignments by a novel graph attention network, which explicitly models the relationships between noun phrases.
- Our GARN achieves the best performance on four challenging benchmarks, which verifies the effectiveness of our model.

The remainder of this paper is organized as follows. In Section II, we introduce related work of image-text matching, identity-aware image-text matching and graph neural network. In Section III, we introduce our GARN model in detail. We present experimental results in Section IV. Finally, we conclude our work in Section V.

II. Related Work

In this section, we introduce the related works, including image-text matching, identity-aware image-text matching and graph neural network.

A. Image-Text Matching

There are many studies exploring mapping the whole image and full sentence to a common feature space for image-text matching [12]–[14]. Kiros *et al.* [12] are the first to learn cross-modal representations with a hinge-based triplet ranking loss, where images are encoded by deep Convolutional Neural Networks (CNN) and textual descriptions are encoded by

Recurrent Neural Networks (RNN) [15]. Faghri *et al.* [13] propose the hard negatives in the triplet loss function and obtain significant gains in retrieval performance. Gu *et al.* [14] propose to incorporate generative models into textual-visual features embedding for cross-modal retrieval. To explore the latent vision-language correspondence between image and text, many works employ the attention mechanism to image-text matching. Huang *et al.* [16] propose a semantic-enhanced image and sentence matching model, which improves the image representation by learning semantic concepts. Lee *et al.* [1] propose a stacked cross attention model to discover the full latent alignments using both image regions and words in sentence as context and infer the image-text similarity. Different from them, we consider to model not only the corresponding relationships between image regions and noun phrases but also the internal relationships between noun phrases and then learn the aligned image-text representations.

B. Identity-Aware Image-Text Matching

Although identity-level annotations are widely used in visual matching tasks, such as person re-identification [17], [18] and face recognition [19]–[21], there are few studies of visual-textual matching. Reed *et al.* [3] collect two datasets with identity-level annotations, namely Caltech-UCSD Birds dataset (CUB) and Oxford-102 Flowers dataset [22]. And they are the first to use identity annotations to learn the image and text features for cross-modal matching. Li *et al.* [10] propose a CUHK Person Description dataset with identity information, which aims to search corresponding person images by the natural language queries. They further employ a CNN-LSTM network with gated neural attention for this task, but they do not effectively utilize the identity-level annotations. To exploit the person identification, Li *et al.* [2] propose an identity-aware two-stage network. First they utilize a Cross-Modal Cross-Entropy loss to embed the input image and description to the same feature space. Then a co-attention mechanism is utilized to refine the network. Zheng *et al.* [23] propose an identification loss for instance-level image-text matching. Zhang and Lu [24] propose a cross-modal projection classification loss to classify the projection of the features from one modality onto the matched features from another modality rather than categorize the original feature representations. These methods show that combining the ranking loss and identification loss can minimize the intra-identity distance and maximize the inter-identity distance simultaneously, so we choose these two loss functions to train our model. In contrast to them, we focus on learning the latent semantic alignment between image and text.

C. Graph Neural Network

Graph neural networks are generally used to handle graph-structured data, which can be divided into two categories. The first class applies Convolutional Neural Networks to graph [25]–[27]. Shen *et al.* [28] create a graph to represent the pairwise relationships between probe-gallery person image pairs (nodes) and utilize such relationships to update the probe-gallery relational features in an end-to-end manner. Yan *et al.* [29] employ the context information for person search and build a graph learning framework to effectively employ context pairs to update the target similarity. Ying *et al.* [30] propose a differentiable graph pooling module that can generate hierarchical representations of graphs and can be combined with various graph neural network architectures in an end-to-end fashion. The second class applies recurrent neural networks to every node of the graph. The messages from the neighbour graph nodes are accumulated and propagated to the nodes, which model the relationships between nodes. There are many studies on the updating of the node hidden state. Scarselli *et al.* [31] propose a multi-layer perceptrons (MLP) to update the hidden state. Gated Graph Neural Network (GGNN) [32] uses gated recurrent units to update the hidden state. Liang *et al.* [33] update the hidden state based on LSTM. Palm *et al.* [34] propose recurrent relational networks in a graph to solve the multi-steps relational reasoning task. Qi *et al.* [35] use 3D graph neural network for semantic segmentation. Si *et al.* [36] use a graph neural network for skeleton-based action recognition. In this paper, we propose a graph attentive relational network to learn the aligned image-text representations.

III. Our Model

In this section, we introduce the graph attentive relational network (GARN) in detail. To learn a more effective textual representation, we propose a skip graph neural network. In addition, we propose a graph attention network to learn the latent semantic alignment between image regions and noun phrases. Besides local matching, the global matching is also employed to learn the global discriminative representations. Finally, we employ a combination of identification loss and pair-wise ranking loss to train the GARN.

A. Visual and Textual Feature Extraction

1) Visual Description: Given an image I, we extract the visual feature using a visual CNN. The image features $\phi'(I) \in \mathbb{R}^{m' \times n \times d}$ are obtained before the last pooling layer of the visual CNN. Then we partition the $\phi'(I)$ into m horizontal stripes. In each stripe, the vectors in same column are averaged into a single column vector. The $\phi'(I)$ is then transformed into $\phi(I) \in \mathbb{R}^{m \times n \times d}$ where $m \times n \times d$ means there are $m \times n$ regions and each region is represented by a d-dimensional vector. The global visual representation $\psi(I) \in \mathbb{R}^d$ is defined as follows:

$$\psi(I) = avgpool(\phi(I)), \tag{1}$$

where $avgpool$ means average pooling along $m \times n$ regions. The local part features $V(I)$ are gained by average pooling the $\phi(I)$ along the column vector, where $V(I) = \{v_1, v_2, \ldots, v_m\}, v_i \in \mathbb{R}^d$. Note that we only utilize the horizontal features for CUHK-PEDES dataset. For the CUB, Flower and Flickr30K datasets, due to the fact that they do not have similar discriminative horizontal features as CUHK-PEDES, we obtain the local part features $V(I)$ by dividing the image features $\phi'(I)$ into $m' \times n$ visual features.

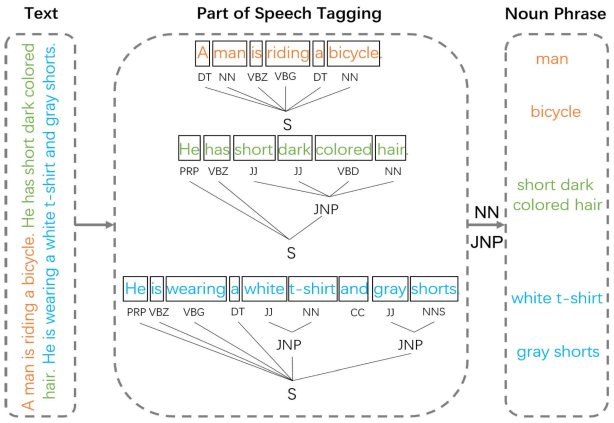

Fig. 3. The illustration of extracting noun phrases from the textual description. We first utilize word-level tokenization and part-of-speech tagging, then extract noun phrases by chunking. Particularly, we extract two kinds of the phrases, e.g., NN and JNP.

2) Textual Description: Given a text T, we first represent every word as a D-dimensional one-hot vector. The j-th word is denoted as $w_j \in \mathbb{R}^D$, where D is the vocabulary size. Then the word is embedded to a p-dimensional vector through an embedding matrix W_e:

$$x_j = W_e w_j, \quad j \in [1, z], \tag{2}$$

where z represents the number of words in text T. Based on the embedding vector, we encode them through a bi-directional long short-term memory network (bi-LSTM) [9] which contains a forward $\overrightarrow{LSTM}$ and a backward $\overleftarrow{LSTM}$:

$$\overrightarrow{h_j} = \overrightarrow{LSTM}(x_j, \overrightarrow{h_{j-1}}), \quad j \in [1, z], \tag{3}$$

$$\overleftarrow{h_j} = \overleftarrow{LSTM}(x_j, \overleftarrow{h_{j-1}}), \quad j \in [1, z]. \tag{4}$$

The LSTM unit inputs the current word embedding vector x_j and previous hidden state h_{j-1}, and outputs the current hidden state h_j.

The global textual representation e^t is defined as the concatenation of the last hidden states $\overrightarrow{h_z}$ and $\overleftarrow{h_1}$:

$$e^t = concat(\overrightarrow{h_z}, \overleftarrow{h_1}). \tag{5}$$

3) Noun Phrase: For the given textual description, we utilize the NLTK [37] to extract the noun phrase N. The extraction procedure is shown in Fig. 3. Similar to textual description, for the j-th noun phrase n_j in $N = (n_1, n_2, \ldots, n_q)$, we represent it according to Equations 2-5. Therefore, we can obtain the representations of all noun phrases $e^n = (e_1^n, e_2^n, \ldots, e_q^n)$. It should be noted that we adopt the same bi-LSTM when encoding the global textual description and noun phrase. Moreover, the number of noun phrase q varies in different textual descriptions.

After obtaining the visual and textual features, the simplest way of measuring the similarity between them is computing the cosine score. But there are some problems as follows. On one hand, directly utilizing the global unaligned features cannot extract the latent correspondences between image regions and noun phrases. On the other hand, the misalignment between textual inputs will compromise the feature learning and matching. We can see it in Fig. 1, the two sentences are both describing the same image but they are different in the describing way. Therefore, learning the aligned image-text representations is of significant value. There are many attention-based methods proposed to solve this problem. They utilize visual (textual) features to focus on textual (visual) features or co-attention. But these attention methods regard different noun phrases in sentences as individuals and ignore the relationships between them which is important to determine whether they belong to the same visual region. For example, the phone is usually held on the hand, so the noun phrases "phone" and "hand" should be divided into the same visual areas of the hand by modeling their relationship. Therefore, the relationships between objects should be modeled for effective matching.

Based on the above analysis, we propose to utilize the graph neural network (GNN) which is excellent to model the relationships between objects to learn aligned image-text representations.

B. Skip-GNN for Textual Representation

First, we use graph neural network to learn effective textual representations. The typical graph is composed of nodes that represent the noun phrases in sentence and edges that represent the relationships between the nodes. Given a set of nodes N and their relationships R, the graph is defined as $G = (N, R)$, where $N = \{n_1, n_2, \ldots, n_q\}$. For the node k in a GNN, the hidden state s_k^t at time step t is updated based on its previous hidden state s_k^{t-1} and message η_k^t received from its neighborhoods Ω_k in a recurrent way. All the nodes are updated simultaneously. Therefore, the formulation of GNN is defined as follows:

$$\eta_k^t = f(\{s_{k'}^{t-1} | k' \in \Omega_k\}), \tag{6}$$

$$s_k^t = g(\eta_k^t, s_k^{t-1}), \tag{7}$$

where f is the message passing function, and g is the node updating function.

In this work, considering that the great success achieved by ResNet and UNet indicates that skip connection is very effective for model optimization and performance improvement, we propose a skip-GNN to model the relationships between noun phrases as well as their initial features encoded by bi-LSTM. The initial features $e^n = (e_1^n, e_2^n, \ldots, e_q^n)$ are fed to skip-GNN as the initial inputs. Fig. 4 shows the structure and updating mechanism of our fully-connected skip-GNN model with four nodes for simplicity. We can see that at time step t, the k-th node inputs an input feature a_k^{t-1} and a message η_k^t. We initialize a_k^0 with the initial feature of noun phrase e_k^n, so that:

$$a_k^0 = W_a e_k^n + b_a, \tag{8}$$

where W_a is the input embedding matrix. Because the node has different relationships with different neighborhood nodes, we utilize the previous hidden state of neighborhood node to define the message. Therefore, the nodes with similar features

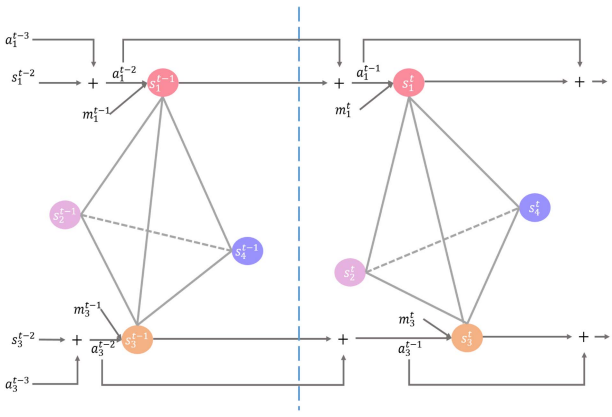

Fig. 4. The illustration of the proposed structure and updating mechanism of the fully-connected skip graph neural network. The nodes s are updated by the passing messages η and input features a in a recurrent way. The skip connection means the connection between a^{t-1} and a^t.

are more closely related to each other.

$$\eta_{k,j}^t = W_m s_j^{t-1} + b_m, \tag{9}$$

$$\eta_k^t = \sum_{j \in \Omega_k} \eta_{k,j}^t, \tag{10}$$

where W_m is the shared message embedding matrix, and η_k^t represents the whole received messages. Then we concatenate the η_k^t and a_k^{t-1} as the final input message. So the nodes can receive the messages from not only their neighbours but also their own initial features. With the obtained message, the hidden state of the node can be updated:

$$s_k^t = g(concat(\eta_k^t, a_k^{t-1}), s_k^{t-1}), \tag{11}$$

where g indicates the node updating function which is similar to the LSTM unit:

$$f_k^t = \sigma(W_f \cdot [s_k^{t-1}, \eta_k^t, a_k^{t-1}] + b_f), \tag{12}$$

$$i_k^t = \sigma(W_i \cdot [s_k^{t-1}, \eta_k^t, a_k^{t-1}] + b_i), \tag{13}$$

$$\tilde{C}_k^t = \tanh(W_C \cdot [s_k^{t-1}, \eta_k^t, a_k^{t-1}] + b_C), \tag{14}$$

$$C_k^t = f_k^t * C_k^{t-1} + i_k^t * \tilde{C}_k^t, \tag{15}$$

$$o_k^t = \sigma(W_o \cdot [s_k^{t-1}, \eta_k^t, a_k^{t-1}] + b_o), \tag{16}$$

$$s_k^t = o_k^t * \tanh(C_k^t), \tag{17}$$

where $W_f, b_f, W_i, b_i, W_C, b_C, W_o, b_o$ are the learned parameters. It's worth noting that these parameters are shared among different nodes.

Then we update the input feature as follows:

$$a_k^t = a_k^{t-1} + s_k^t. \tag{18}$$

As the node features are updated after every time step, this input feature can fuse the initial textual features with node relational features by skip connection between a_k^t and a_k^{t-1}. After iterating the message passing for T steps, we compute the final fusion representations as:

$$p_k = W_p a_k^T + b_p, \tag{19}$$

where W_p is an output embedding matrix.

Due to the fact that the number of noun phrase q varies in different textual descriptions, for textual description with fewer noun phrases than the number of nodes in skip-GNN, we set the hidden states, input messages, and output messages of all unused nodes to zero at every time step to make sure that they cannot receive or send any information.

C. Graph Attention Network for Image-Text Alignment

The relationships between noun phrases indicate whether they belong to the same visual region. To obtain the probabilities that the noun phrases belong to the image regions, we propose a graph attention network by modeling the relationships between noun phrases.

The graph attention network aims to learn the attention matrix over the nodes of skip-GNN model, which can extract the node embeddings that are corresponding to specific image regions. We first describe the generation of the attention matrix using a GNN architecture and then discuss attention procedure given the attention matrix.

We generate the attention matrix by a typical GNN as follows:

$$A = softmax(GNN_{att}(e^n)), \tag{20}$$

where GNN_{att} means the same operation as Equations 6 and 7, the softmax function is applied in a row-wise fashion. e^n is the noun phrase features, which is fed to the typical GNN. The output dimension of GNN_{att} corresponds to the pre-defined number m. Therefore, the attention matrix $A \in \mathbb{R}^{q \times m}$.

With the attention matrix, we discuss the attention procedure. Each row of A corresponds to one of the q noun phrase representations from skip-GNN and each column of A corresponds to one of the image regions, which provides a soft assignment of each noun phrase representation to the image regions. With the computed A, we perform the following operation:

$$Att = A^T P, \quad Att \in \mathbb{R}^{m \times 2l}, \tag{21}$$

where attention matrix A aggregates the noun phrase representations $P = \{p_1, p_2, \ldots, p_q\}$ to the part-level cluster, l is the hidden dimension of the bi-LSTM in textual representation learning.

D. Local and Global Matching

With the learned aligned image-text representations, we measure the local similarity between image regions and noun phrases. First, we transform the image region features $V(I)$ and attended noun phrase representations Att into the same feature space:

$$\tilde{v_i} = W_v v^i, \tag{22}$$

$$\widetilde{att_i} = W_{att} att_i, \quad i = 1, 2, \ldots, m, \tag{23}$$

where $W_v \in \mathbb{R}^{b \times d}$ and $W_{att} \in \mathbb{R}^{b \times 2l}$ are two transformation matrices, and b is the dimension of the transformed feature space. The att_i represents the i-th row vector of Att.

Then, the local similarity between image region and noun phrase is defined as:

$$s_i = cos(\tilde{v_i}, \tilde{at_i}), \quad i = 1, 2, \ldots, m, \tag{24}$$

$$S^l = \sum_{i=1}^{m} s_i, \tag{25}$$

where cos represents the cosine function.

Besides local matching, the global matching is also utilized to measure their global similarity. We calculate the global correlation between global visual representation $\psi(I)$ and textual representation e^t.

We first transform global visual representation $\psi(I)$ and textual representation e^t to the same feature space as follows:

$$\tilde{e^t} = W_{e^t} e^t, \tag{26}$$

$$\widetilde{\psi(I)} = W_\psi \psi(I), \tag{27}$$

where $W_{e^t} \in \mathbb{R}^{b \times 2l}$ and $W_\psi \in \mathbb{R}^{b \times d}$ are two transformation matrices.

The global similarity is then computed as follows:

$$S^g = \cos(\widetilde{\psi(I)}, \tilde{e^t}). \tag{28}$$

E. Learning Procedure

The pair-wise ranking loss is the common loss function used in the matching task, which aims to ensure the positive pair being closer than the negative pair. Many previous works randomly select the negative pair from the dataset and ignore the influence of other negative samples in a mini-batch. In this paper, we follow the method [13] to focus the hardest negative sample in a mini-batch. Given a positive pair (I_p, T_p), the hardest negative pair is defined as follows:

$$T_{\widehat{h}} = argmax_{t \neq T} S(I, t), \tag{29}$$

$$I_{\widehat{h}} = argmax_{i \neq I} S(i, T), \tag{30}$$

where $T_{\widehat{h}}$ is the hardest text sample for the image I_p and $I_{\widehat{h}}$ is the hardest image sample for the text T_p. Therefore, our ranking loss is defined as:

$$\begin{aligned} L_r(I, T) = & max(\alpha - S(I, T) + S(I, T_{\widehat{h}}), 0) \\ & + max(\alpha - S(I, T) + S(I_{\widehat{h}}, T), 0), \end{aligned} \tag{31}$$

where α is a margin. This loss function ensures the positive pair being closer than the hardest negative pair which may determine success or failure as measured by top-1 accuracy. For our global matching score, we can obtain a global ranking loss L_r^g.

Besides ranking loss, the identification loss is also adopted for the identity-level matching. The global image and text identification losses L_i^g and L_t^g are defined as follows:

$$L_i^g = -y_{id} log(softmax(W_{id}^g \widetilde{\psi(I)})), \tag{32}$$

$$L_t^g = -y_{id} log(softmax(W_{id}^g \tilde{e^t})), \tag{33}$$

where W_{id} is the transformation matrix to categorize the feature representations, y_{id} is the ground truth identity, L_i^g and L_t^g are the global visual and textual identification losses, respectively.

Then the total global loss is defined as:

$$L^g = L_r^g + \lambda_1 L_i^g + \lambda_2 L_t^g, \tag{34}$$

Similarly, we can obtain the total local loss L^l. λ is the hyperparameter to control the relative importance of each loss function.

The final loss function is defined as:

$$L = L^g + \lambda_3 L^l. \tag{35}$$

At the test stage, we compute the total similarity S between the image-text pair for retrieval evaluation, which is defined as follows:

$$S = S^g + \lambda_3 S^l. \tag{36}$$

IV. Experiments

In this section, we first introduce the experimental datasets. Then, we present the implementation details. Next, we compare the proposed method with the state-of-the-art methods and several baselines. Finally, we visualize and analyze the retrieval results.

A. Datasets

We choose four identity-aware cross-modal matching benchmark datasets to validate the effectiveness of the proposed GARN.

1) CUHK-PEDES Dataset: The CUHK-PEDES dataset [10] is collected from five existing person re-identification datasets, CUHK03 [38], Market-1501 [39], SSM [40], VIPER [41], and CUHK01 [42], as the subjects for language descriptions. All the images were labeled by crowd workers from Amazon Mechanical Turk (AMT). As a result, the CUHK-PEDES dataset contains 40,206 images and 80,440 textual descriptions of 13,003 identities. We follow the same data split as [10]. The training set has 11,003 persons, 34,054 images and 68,126 textual descriptions. The validation set has 1,000 persons, 3,078 images and 6,158 textual descriptions. The test set has 1,000 persons, 3,074 images and 6,156 textual descriptions. On average, each image contains 2 different textual descriptions and the textual descriptions contain more than 23 words. We choose top-1, top-5 and top-10 accuracies to evaluate the performance of person search with natural language description. Specifically, given a query text, all test images are ranked by the similarities with the text. If the corresponding images are within the top-k images, we regard it as a successful search.

2) CUB and Flower Datasets: The Caltech-UCSD Birds (CUB) [3] dataset contains 11,788 bird images, which are categorized into 200 classes. Each image is described by ten sentences. The dataset is split into 100 categories for training, 50 categories for validation, and 50 categories for test. On average, the textual descriptions contain more than 17 words. The Oxford-102 Flowers [3] dataset has 8,189 flower images categorized into 102 classes. Each image is also described by ten sentences. There are 62 categories for training, 20 categories for validation, and 20 categories for test. On average, the textual descriptions contain more

TABLE I
SUMMARY STATISTICS OF FOUR DATASETS (CUHK-PEDES, CUB, FLOWER AND FLICKR30K)

Datasets	Images	Languages	Length	ID	Train	Val	Test	Image/ID	Languge/Image
CUHK-PEDES	40,206	80,440	23	13,003	11,003	1,000	1,000	3	2
CUB	11,788	117,880	17	200	100	50	50	59	10
Flower	8,189	81,890	14	102	62	20	20	80	10

than 14 words. The experimental setup is the same as [3]. We choose AP@50 to evaluate the text-to-image retrieval performance and top-1 accuracy to evaluate the image-to-text retrieval performance. The AP@50 represents the percent of top-50 ranked images whose class matches that of the text query, averaged over all the test classes.

3) Flickr30K Dataset: The Flickr30K [11] contains 31,783 images and each image is annotated with five descriptions. The average sentence length is 11 words. We follow the same split in [13] to use 29,783 images for training, 1,000 images for validation and 1,000 images for testing.

Table I shows the summary statistics of these four datasets.

B. Implementation Details

For the CUHK-PEDES dataset, we perform experiments with VGG-16 [43], ResNet-50 [44] and MobileNet [45] as the visual CNN. The input images are resized to 384×128, with a height to width ratio of 3:1. We set the visual features to $m = 6$ horizontal stripes inspired by [46]. To compare with previous methods fairly, we choose GoogleNet [47] as the visual CNN for the CUB and Flower datasets. The images are resized to 299×299. For the Flickr30K dataset, we choose ResNet-152 [44] as the visual CNN. The images are resized to 224×224. Due to the fact that the CUB, Flower and Flickr30K datasets do not have similar discriminative horizontal features as CUHK-PEDES, we use the image features $\phi'(I)$ to compute the local similarity. For all the four datasets, a $l = 1024$ dimensional bi-LSTM is used to extract the textual feature. We embed the word to a $p = 300$ dimensional vector and set the dimension b of the transformed feature space as 1024. In graph neural network, the iteration steps T in skip graph neural network and graph attention network are set to 3 and 2, respectively.

During training, we first fix the visual CNN and train the other parts with learning rate $lr = 2e^{-3}$, and then train the whole model with learning rate $lr = 2e^{-4}$. The Adam optimizer [48] is employed for optimization and the margin is set to 0.2. The identity classes are only used for training.

To exploit the influences of λ_1, λ_2 and λ_3, we set them to $\{0.1, 0.2, 0.5, 0.7, 1, 2, 5, 7, 10\}$ and find that the model with $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 2$ obtains the best performance on the validation set.

C. Experimental Results

1) Results on the CUHK-PEDES Dataset: We compare our proposed GARN with the state-of-the-art methods on the CUHK-PEDES dataset. Table II shows the results of top-1, top-5 and top-10 accuracies with three different visual CNNs (VGG-16, ResNet-50, Mobilenet). Considering that this dataset is designed to search the corresponding person images to the textual description, we only show the text-to-image retrieval results. Overall, it can be seen that the proposed GARN achieves the best performances in terms of VGG-16, ResNet-50 and Mobilenet. Specifically, when comparing with the best competitor Dual Path [23] using VGG-16 and ResNet-50 to extract visual representation, our GARN significantly outperforms it by about 14% with the VGG-16 feature and 8% with the ResNet-50 feature, respectively. The improved performances over the best competitor indicate that our GARN is very effective for this task. Although CMPM+CMPC [24] achieves the best result (49.37%) by virtue of Mobilenet, our GARN with the same setting still improves the performance by 3.4% in top-1 accuracy. Compared with the methods (PWM-ATH [49], GNA-RNN [10], GLA [50] and IATV [2]) which aim to align image and text representations by either utilizing the textual representation to focus on the visual unit or employing a co-attention to select both visual and textual representations, our GARN also achieves better performances under three evaluation metrics. The improved performances illustrate the superiorities of our graph attentive relational network in learning aligned image-text representations by modeling the relationship between noun phrases.

TABLE II
COMPARISON RESULTS WITH THE STATE-OF-THE-ART METHODS ON CUHK-PEDES. TOP-1, TOP-5 AND TOP-10 ACCURACIES (%) OF TEXT-TO-IMAGE RETRIEVAL RESULTS ARE REPORTED. THREE TYPES OF VISUAL CNN ARE UTILIZED, E.G., VGG-16, RESNET-50, AND MOBILENET. THE BEST PERFORMANCE IS **BOLD**. "-" REPRESENTS THAT THE RESULT IS NOT PROVIDED

Method	Visual	Top-1	Top-5	Top-10
LSTM Q+norm I[4]	VGG-16	17.19	-	57.82
CNN-RNN[3]	VGG-16	8.07	-	32.47
Neural Talk[50]	VGG-16	13.66	-	41.72
GNA-RNN[10]	VGG-16	19.05	-	53.64
IATV[2]	VGG-16	25.94	-	60.48
PWM-ATH[48]	VGG-16	27.14	49.45	61.02
Dual Path[22]	VGG-16	32.15	54.42	64.30
GARN(ours)	**VGG-16**	**46.25**	**67.48**	**76.84**
GLA[49]	Res-50	43.58	66.93	76.26
Dual Path[22]	Res-50	44.40	66.26	75.07
GARN(ours)	**Res-50**	**52.25**	**73.51**	**81.12**
CMPM+CMPC[23]	Mobilenet	49.37	-	79.27
GARN(ours)	**Mobilenet**	**52.75**	**74.36**	**81.85**

2) Results on the CUB and Flower Datasets: Table III and table IV show the retrieval results on the CUB and Flower datasets, respectively. Considering that we use the bi-directional losses in our experiments, we choose the symmetric results of the existing methods for fair comparison. It should be noted that except CMPM+CMPC with

TABLE III

COMPARISON RESULTS WITH THE STATE-OF-THE-ART METHODS ON CUB DATASET. THE ACCURACY (%) OF IMAGE-TO-TEXT (TOP-1) AND TEXT-TO-IMAGE (AP@50) RETRIEVAL RESULTS ARE REPORTED. THE BEST PERFORMANCE IS **BOLD**

Method	Image-Text Top-1	Text-Image AP@50
BoW[51]	44.1	39.6
Word2Vec[52]	38.6	33.5
Word CNN[3]	51.0	43.3
Word CNN-RNN[3]	56.8	48.7
GMM+HGLMM[53]	36.5	35.6
Triplet[2]	52.5	52.4
IATV[2]	61.5	57.6
ABM[54]	60.2	48.3
CMPM+CMPC[23]	64.3	67.9
GARN(ours)	**69.7**	**69.4**

TABLE IV

COMPARISON RESULTS WITH THE STATE-OF-THE-ART METHODS ON FLOWER DATASET. THE ACCURACY (%) OF IMAGE-TO-TEXT (TOP-1) AND TEXT-TO-IMAGE (AP@50) RETRIEVAL RESULTS ARE REPORTED. THE BEST PERFORMANCE IS **BOLD**

Method	Image-Text Top-1	Text-Image AP@50
BoW[51]	57.7	57.3
Word2Vec[52]	54.2	52.1
Word CNN[3]	60.7	56.3
Word CNN-RNN[3]	65.6	59.6
GMM+HGLMM[53]	54.8	52.8
Triplet[2]	64.3	64.9
IATV[2]	68.4	70.1
ABM[54]	68.7	60.2
CMPM+CMPC[23]	68.9	69.7
GARN(ours)	**71.8**	**72.4**

Mobilenet in [24], the other methods utilize GooleNet as visual CNN. The BoW [52], Word2Vec [53], Word CNN [3], Word CNN-RNN [3], and GMM+HGLMM [54] use different types of textual representations. For our GARN, a bi-LSTM is used to represent the textual input. The Triplet [2] employs a triplet loss to train the model, but the ABM [55] proposes an angle-based loss function. CMPM+CMPC [24] employs a combination of Cross-Modal Projection Matching (CMPM) loss and Cross-Modal Projection Classification (CMPC) loss to train the model while the IATV [2] proposes a co-attention approach. Different from them, we propose a graph attentive relational network to learn the aligned features and use the ranking loss and identification loss to train the model. We can see that our GARN achieves the state-of-the-art performances on both CUB and Flower datasets, which are 69.7%, 71.8% in top-1 accuracy for image-to-text retrieval and 69.4%, 72.4% in AP@50 for text-to-image retrieval, respectively. This proves the effectiveness of our proposed graph-based aligned feature learning in identity-aware cross-modal matching task.

3) Results on the Flickr30K Dataset: Table V shows the retrieval results on the Flickr30K dataset. For fair comparison, we choose the existing methods (e.g., RRF-Net [56], VSE++ [13], DAN [57], DPC [23] and SCO [16]) which utilize the ResNet-152 as the visual backbone and do not utilize the Faster-RCNN to detect objects like us. We can see that the proposed GARN outperforms the previous methods, which demonstrates the effectiveness of graph-based representations learning for image-text matching.

TABLE V

COMPARISON RESULTS WITH THE STATE-OF-THE-ART METHODS ON FLICKR30K DATASET. TOP-1, TOP-5 AND TOP-10 ACCURACIES (%) OF BI-DIRECTIONAL RETRIEVAL RESULTS ARE REPORTED. THE BEST PERFORMANCE IS **BOLD**

Method	Image-to-Text			Text-to-Image		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
RRF-Net[56]	47.6	77.4	87.1	35.4	68.3	79.9
VSE++[13]	52.9	80.5	87.2	39.6	70.1	79.5
DAN[57]	55.0	81.8	89.0	39.4	69.2	79.1
DPC[23]	55.6	81.9	89.5	39.1	69.2	80.9
SCO[16]	55.5	82.0	89.3	41.1	70.5	**81.1**
GARN(ours)	**60.1**	**84.6**	**90.4**	**44.2**	**71.2**	80.3

TABLE VI

ABLATION ANALYSIS OF DIFFERENT COMPONENTS IN THE PROPOSED GARN ON CUHK-PEDES. L_{id} INDICATES THE IDENTIFICATION LOSS. SGNN REPRESENTS THE SKIP GRAPH NEURAL NETWORK IN LEARNING EFFECTIVE TEXTUAL REPRESENTATION. G-ATTENTION INDICATES THE GRAPH ATTENTION NETWORK IN LEARNING THE LATENT SEMANTIC ALIGNMENT. GLOBAL INDICATES THAT THE GLOBAL FEATURES ARE USED IN CROSS-MODAL MATCHING. S-ATTENTION MEANS THE SIMILARITY-BASED ATTENTION. SINGLE-GRAPH INDICATES UTILIZING A GRAPH NEURAL NETWORK TO PERFORM THE GRAPH ATTENTIVE RELATIONAL LEARNING. RESNET-50 IS UTILIZED AS THE VISUAL CNN. TOP-1, TOP-5 AND TOP-10 ACCURACIES (%) ARE REPORTED

Method	Top-1	Top-5	Top-10
Base	47.32	68.75	78.91
Base+L_{id}	48.67	70.53	79.22
Base+L_{id}+SGNN	50.24	71.62	80.06
Base+L_{id}+S-Attention	49.50	70.42	79.79
Base+L_{id}+G-Attention	50.78	71.11	80.24
Base+L_{id}+SGNN+G-Attention	51.84	72.69	81.02
Base+L_{id}+Single-Graph	51.04	72.20	80.88
Base+L_{id}+SGNN+G-Attention+Global	52.25	73.51	81.12

D. Model Analysis

1) Ablation Studies: To systematically investigate the effectiveness of each component in the proposed GARN, we perform a set of ablation studies on the CUHK-PEDES dataset. It's worth noting that we utilize ResNet-50 as the visual CNN. Table VI shows the results. To make a better comparison, we set a baseline model named Base, which employs a ResNet-50 to extract the visual feature and a same bi-LSTM as GARN to encode the textual input. Then an embedding layer is utilized to transform the cross-modal features into the same feature space. Only the ranking loss L_r is used to train the model.

We first investigate the importance of identification loss by adding the visual and textual identification losses L_i and L_t to the baseline model, which is denoted as Base+L_{id}. It can be seen that the top-1 accuracy rises 1.3% compared with Base, which proves the effectiveness of identification loss in

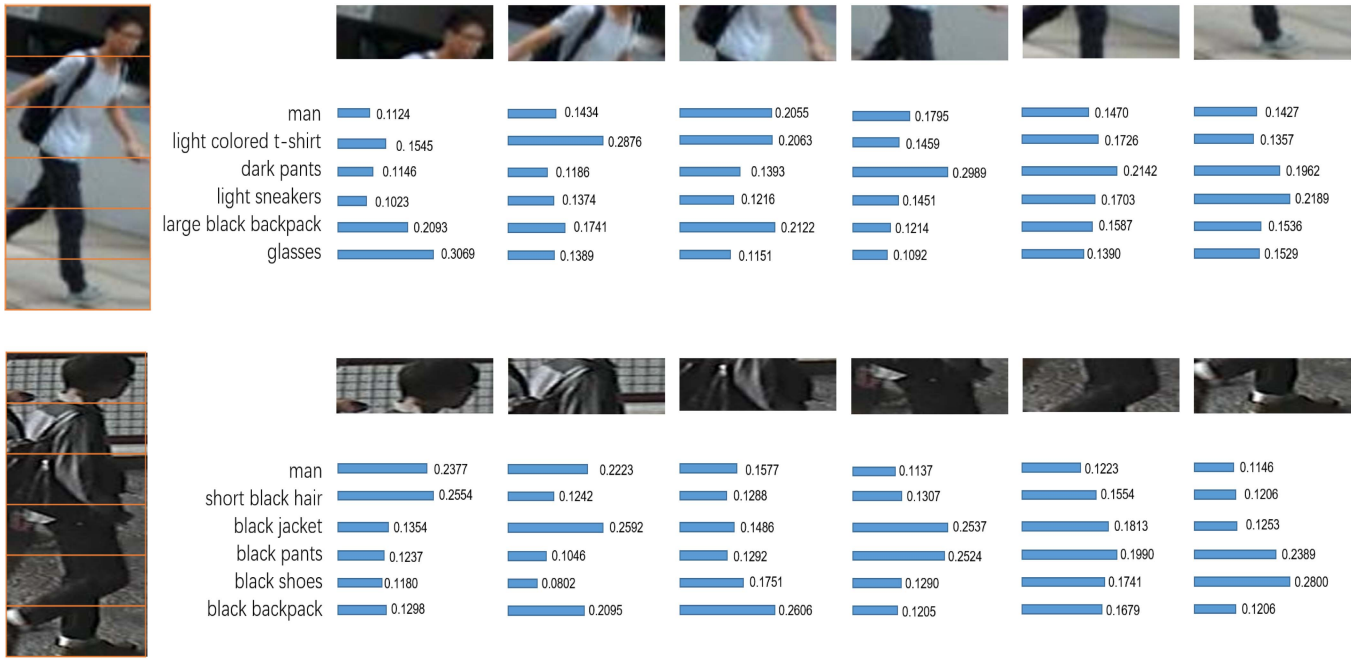

Fig. 5. Visualization of the attended noun phrases with respect to each image region on two examples from CUHK-PEDES by our proposed GARN. The value indicates the attention strength.

benefitting identity-level matching. Then we investigate the effectiveness of skip graph neural network in learning effective textual representation by adding a skip graph neural network into the Base+L_{id}, which is denoted as Base+L_{id}+SGNN. The top-1 accuracy rises 1.6% compared with Base+L_{id}, which indicates that SGNN can help our model learn more discriminative textual representation by modeling the relationships between noun phrases and thus benefit the performances. In addition, we perform the experiment with general graph neural network to learn the textual representation but do not obtain meaningful results. This is because the hardest ranking loss in general GNN model is difficult to decrease if not elaborating the initialization of the weight parameters. Considering that the skip connection is very effective for model optimization and performance improvement (e.g., ResNet and UNet), we add skip connection in our model. To investigate the importance of graph attention network in learning latent semantic alignment between visual and textual representations, we perform experiments on Base+L_{id}+G-Attention and Base+L_{id}+SGNN+G-Attention, where G-Attention indicates the graph attention network. The improved performances prove that graph attention network benefits the aligned cross-modal representations learning. We also perform another attention mechanism namely similarity-based attention (S-Attention) proposed in [1]. The results are inferior to the G-Attention by 1.2% in top-1 accuracy. Due to the fact that the two graph neural networks in Base+L_{id}+SGNN+G-Attention are different, we also perform the experiment utilizing the same graph neural network, which is denoted as Base+L_{id}+Single-Graph. The results indicate that it is more appropriate to use two graph neural networks to learn attention matrix and textual representations, respectively. Since the above learning is based on phrase-level matching, to exploit different levels cross-modal matching, we add the global matching into the Base+L_{id}+SGNN+G-Attention, which is denoted as Base+L_{id}+SGNN+G-Attention+Global. It can be seen that the top-1 accuracy rises 0.4% compared with Base+L_{id}+SGNN+G-Attention, which proves that exploiting different levels cross-modal matching is effective in image-text matching by learning sufficient and diverse discriminative representations. In summary, the improved performances demonstrate that identification loss, skip graph neural network, graph attention network and different level features are all effective for identity-aware image-text matching.

2) Message Propagation Analysis: The number of message propagation T is an important hyperparameter, which determines the information transmitted between the noun phrases. From the results reported in Table VII, we can see that increasing T in skip graph neural network improves the prediction performance and saturates soon. This is because noun phrases are fully connected to each other, so the relationships between them are learned quickly. Considering the performance and running speed, we choose $T = 3$ in our experiments.

3) Qualitative Results: To verify whether the proposed GARN can learn the aligned image-text representations by the graph attentive relational network and make matching procedure more interpretable, we visualize the attended noun phrases with respect to each image region on CUHK-PEDES in Fig. 5. For the two selected images, we split them horizontally into six regions, and visualize the attention weights to each noun phrase in textual descriptions "The man has on a light colored t-shirt with dark pants, and light sneakers. he has a large black backpack and glasses." and "A man with short black hair is wearing a black jacket, a pair of black pants, black shoes and a black backpack". We can see that the selected noun phrases are indeed corresponding to the image regions,

Fig. 6. Qualitative results of image retrieval given text queries on CUHK-PEDES dataset by two models (baseline and GARN). We show the top-10 retrieved images for each query, which are sorted by their similarity scores with text. We outline the corresponding images in green boxes and unmatched images in red boxes.

TABLE VII
THE COMPARISON RESULTS ON CUHK-PEDES DATASET IN ACCURACY (%). WE COMPARE SEVERAL MODELS THAT HAVE DIFFERENT TIME STEPS IN SKIP GRAPH NEURAL NETWORK TO SHOW THE IMPROVEMENTS ACHIEVED AT EVERY STEP

GARN	Top-1	Top-5	Top-10
T=1	50.94	71.68	78.95
T=2	51.89	72.34	80.86
T=3	52.25	73.51	81.12
T=4	52.04	73.45	80.94
T=5	51.94	73.69	81.04

which proves that our model can learn accurate latent semantic alignment between image regions and noun phrases by the graph attention mechanism. Specifically, in the first image, the "glasses" receives strong attention with respect to the first image region which is mainly about the head of person. For the third region of the first image, the "man", "light colored t-shirt" and "large black backpack" receive strong attention while "glasses" receives weaker attention weight. Similarly, for the second image, our proposed GARN can also learn the correspondence between image region and noun phrase. This illustrates that our graph attention network learns interpretable aligned cross-modal representations and generates reasonable attention strength to weight noun phrases, which benefits the inference of image-text similarity.

To better show the retrieval results of proposed GARN, we perform several qualitative evaluations. Fig. 6 shows the qualitative results of image retrieval given text queries on CUHK-PEDES by two models (baseline and GARN). We show the top-10 images which are ranked by the similarity scores with text queries. The matched images are outlined in green while the unmatched images are outlined in red. As Fig. 6 shows, for all the four samples, our GARN retrieves the corresponding images in the top-10 list. For the two cases in the first row, the first three results retrieved by both models are corresponding to the text query. This shows that both models can learn discriminative features for easily distinguishable samples. For the cases in the second and third rows, our GARN achieves better performances for more difficult samples. This demonstrates the effectiveness of our GARN in identity-aware cross-modal matching. In addition, all the images with the same identity as text appear in the top-10 retrieved images, which demonstrates the effectiveness

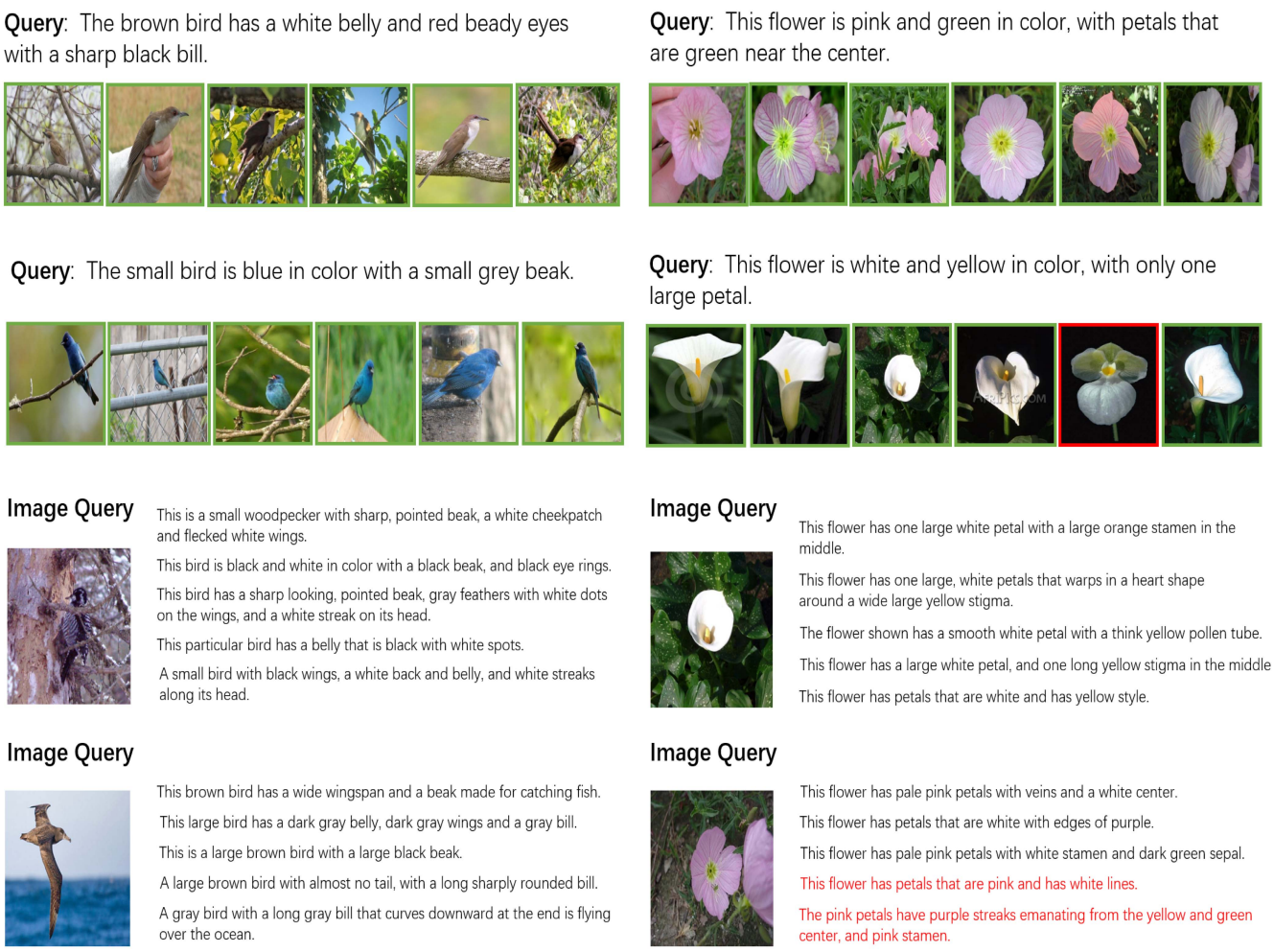

Fig. 7. Qualitative results of image retrieval given text queries and text retrieval given image queries on CUB and Flower datasets. We show the top-6 retrieved images and top-5 retrieved texts for each query, which are sorted by their similarity scores with query. We outline the corresponding images in green boxes and unmatched images in red boxes. And the unmatched texts are highlighted in red.

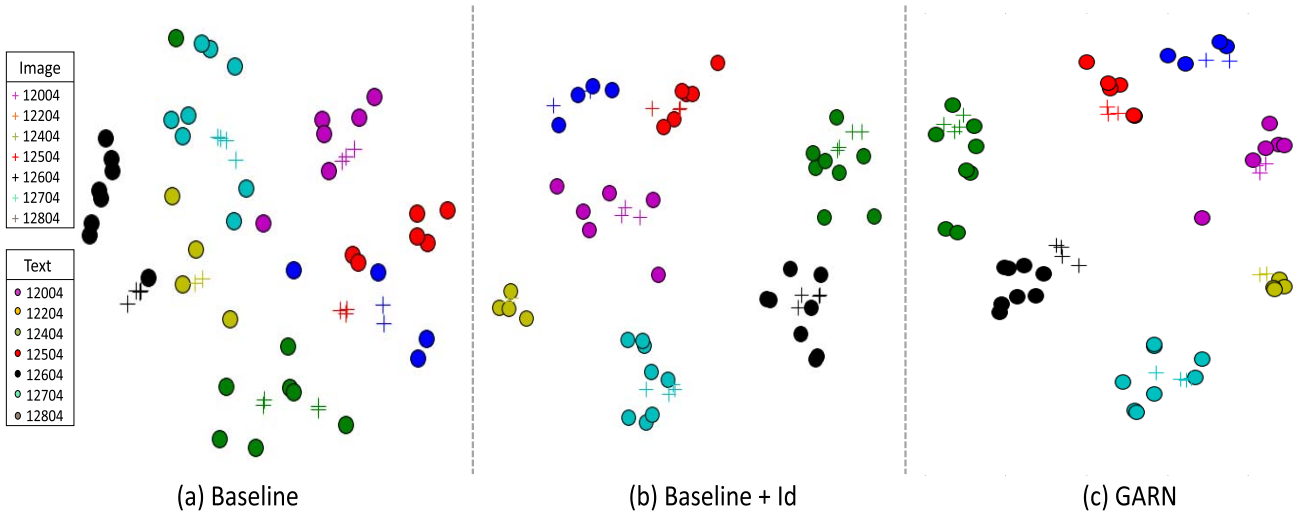

Fig. 8. The t-SNE visualization of image and text features learned by our proposed baseline, baseline+id and GARN on CUHK-PEDES dataset.

of our model in matching the visual and textual inputs of the same identity. For the cases in the forth row, both models capture the part matching, e.g., the person almost wear a "light colored shirt" and "black slacks". But for indistinguishable

details "long light colored top", our GARN can still capture it and retrieve the correct result.

Similarly, we visualize the retrieval results of proposed GARN on CUB and Flower. Fig. 7 shows the qualitative results of image retrieval given text queries and text retrieval given image queries. We show the top-6 retrieved images and top-5 retrieved texts for each query, which are sorted by their similarity scores with query. We outline the corresponding images in green boxes and unmatched images in red boxes. And the unmatched texts are highlighted in red. For the text query cases in the first two rows, the top-6 ranked images almost are corresponding to textual descriptions. The only failure case mismatches the "only one large petal". And for the first three image query cases, the top-5 retrieved texts are corresponding to images. The failure case mismatches the center in flower. It's worth noting that there are many more images with the same identity in the CUB and Flower datasets than in the CUHK-PEDES. Therefore, the visualization results on CUB and Flower appear better.

We also visualize the distribution of cross-modal features learned by our models (baseline, baseline+id and GARN) on CUHK-PEDES dataset, which aims to figure out whether our model can match the corresponding cross-modal features well. Fig. 8 shows the t-SNE [58] visualization of the test feature distribution. We randomly select seven identities and show all corresponding images and texts. We can see that the learned image-text features from all models are distributed along radial spokes, where the corresponding visual and textual features lie in the same direction. This is because we utilize the cosine score to measure the similarity. Compared with non-corresponding pairs, the corresponding image-text pairs have larger cosine score. When adding the identification loss to minimize the intra-identity distance, the learned features are more aggregated with the same identity. Compared to baseline+id model, the learned features by GARN with different identities are more differentiated along radial spokes, which proves the effectiveness of our GARN in learning more discriminative features by alignment.

V. Conclusion

In this paper, we propose a graph attentive relational network to learn the aligned image-text representations for the identity-aware image-text matching. Our main contributions are improving the textual representation and learning the semantic alignment between image and text by modeling the relationships between noun phrases. These are accomplished by skip graph neural network and graph attention network, respectively. In the matching procedure, both the global matching and local matching are utilized to learn more discriminative representations. We perform extensive experiments on four identity-aware datasets, and the experimental results show that our approach achieves much better performances than the state-of-the-art methods, which verifies the effectiveness of our GARN in identity-aware image-text matching.

References

[1] 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.*, Sep. 2018, pp. 201–216.

[2] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1890–1899.

[3] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.

[4] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[5] H. Xue, W. Chu, Z. Zhao, and D. Cai, "A better way to attend: Attention with trees for video question answering," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5563–5574, Nov. 2018.

[6] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[7] S. Ye, J. Han, and N. Liu, "Attentive linear transformation for image captioning," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5514–5524, Nov. 2018.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1970–1979.

[11] 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, Dec. 2014.

[12] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*. [Online]. Available: http://arxiv.org/abs/1411.2539

[13] 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.

[14] 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.

[15] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010.

[16] 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.

[17] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4099–4108.

[18] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.

[19] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4838–4846.

[20] M. Kafai, L. An, and B. Bhanu, "Reference face graph for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2132–2143, Dec. 2014.

[21] L. He, H. Li, Q. Zhang, and Z. Sun, "Dynamic feature matching for partial face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 791–802, Feb. 2019.

[22] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[23] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen, "Dual-path convolutional image-text embedding with instance loss," 2017, *arXiv:1711.05535*. [Online]. Available: http://arxiv.org/abs/1711.05535

[24] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 686–701.

[25] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[26] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3844–3852.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.
[28] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 486–504.
[29] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2158–2167.
[30] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4800–4810.
[31] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.
[32] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," 2015, *arXiv:1511.05493*. [Online]. Available: http://arxiv.org/abs/1511.05493
[33] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph LSTM," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 125–143.
[34] R. Berg Palm, U. Paquet, and O. Winther, "Recurrent relational networks," 2017, *arXiv:1711.08028*. [Online]. Available: http://arxiv.org/abs/1711.08028
[35] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3D graph neural networks for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5199–5208.
[36] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 103–118.
[37] E. Loper and S. Bird, "NLTK: The natural language toolkit," 2002, *arXiv:cs/0205028*. [Online]. Available: https://arxiv.org/abs/cs/0205028
[38] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian, "Person re-identification meets image search," 2015, *arXiv:1502.02171*. [Online]. Available: http://arxiv.org/abs/1502.02171
[40] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," 2016, *arXiv:1604.01850v1*. [Online]. Available: https://arxiv.org/abs/1604.01850v1
[41] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill.*, 2007, pp. 1–7.
[42] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 31–44.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[45] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[46] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 480–496.
[47] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[49] T. Chen, C. Xu, and J. Luo, "Improving text-based person search by spatial matching and adaptive threshold," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1879–1887.
[50] D. Chen, H. Li, X. Liu, Y. Shen, Z. Yuan, and X. Wang, "Improving deep visual representation for person re-identification by global and local image-language association," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 54–70.
[51] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
[52] Z. S. Harris, "Distributional structure," *Word*, vol. 10, nos. 2–3, pp. 146–162, Aug. 1954.
[53] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[54] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4437–4446.
[55] C. Tang, J. Lv, Y. Chen, and J. Guo, "An angle-based method for measuring the semantic similarity between visual and textual features," *Soft Comput.*, vol. 23, pp. 4041–4050, Feb. 2018.
[56] 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.
[57] 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.
[58] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, 2014.

Ya Jing received the B.S. degree from the Department of Automation, Beihang University, in 2016. She is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. She has published related papers in the leading international conferences, such as CVPR and AAAI. Her research interests include situation recognition, text-based person retrieval, action recognition, and deep learning.

Wei Wang received the B.E. degree from the Department of Automation, Wuhan University, in 2005, and the Ph.D. degree from the School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences (GUCAS), in 2011. He is currently an Associate Professor with Center for Research on Intelligent Perception and Computing and the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He has published a number of papers in the leading international conferences, such as CVPR and ICCV. His research interests include computer vision, pattern recognition, and machine learning, particularly on the computational modeling of visual attention, deep learning, and multimodal data analysis.

Liang Wang (Fellow, IEEE) received the B.S. and M.S. degrees from Anhui University in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he has been working as a Research Assistant with Imperial College London, U.K., and Monash University, Australia, a Research Fellow with The University of Melbourne, Australia, and a Lecturer with the University of Bath, U.K. He is currently a Full Professor of Hundred Talents Program at the National Laboratory of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published at highly-ranked international journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON IMAGE PROCESSING, and leading international conferences, such as CVPR, ICCV, and ICDM. He is currently a Fellow of IAPR. He is an Associate Editor of IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: CYBERNETICS.

Tieniu Tan (Fellow, IEEE) received the B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and the M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. He is currently a Professor with Center for Research on Intelligent Perception and Computing, NLPR, CASIA, China. He has published more than 450 research papers in refereed international journals and conferences in the areas of image processing, computer vision, and pattern recognition, and has authored or edited 11 books. His research interests include biometrics, image and video understanding, information hiding, and information forensics. He is a Fellow of CAS, TWAS, BAS, IAPR, and the U.K. Royal Academy of Engineering, and the Past President of IEEE Biometrics Council.