Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search

Jiawei Liu
University of Science and Technology of China
ljw368@mail.ustc.edu.cn

Zheng-Jun Zha*
University of Science and Technology of China
zhazj@ustc.edu.cn

Richang Hong
Hefei University of Technology
hongrc@hfut.edu.cn

Meng Wang
Hefei University of Technology
eric.mengwang@gmail.com

Yongdong Zhang
University of Science and Technology of China
zhyd73@ustc.edu.cn

ABSTRACT

The newly emerging text-based person search task aims at retrieving the target pedestrian by a query in natural language with fine-grained description of a pedestrian. It is more applicable in reality without the requirement of image/video query of a pedestrian, as compared to image/video based person search, *i.e.*, person re-identification. In this work, we propose a novel deep adversarial graph attention convolution network (A-GANet) for text-based person search. The A-GANet exploits both textual and visual scene graphs, consisting of object properties and relationships, from the text queries and gallery images of pedestrians, towards learning informative textual and visual representations. It learns an effective joint textual-visual latent feature space in adversarial learning manner, bridging modality gap and facilitating pedestrian matching. Specifically, the A-GANet consists of an image graph attention network, a text graph attention network and an adversarial learning module. The image and text graph attention networks are designed with a novel graph attention convolution layer, which effectively exploits graph structure in the learning of textual and visual features, leading to precise and discriminative representations. An adversarial learning module is developed with a feature transformer and a modality discriminator, to learn a joint textual-visual feature space for cross-modality matching. Extensive experimental results on two challenging benchmarks, *i.e.*, CUHK-PEDES and Flickr30k datasets, have demonstrated the effectiveness of the proposed method.

CCS CONCEPTS

• **Information systems** → **Top-k retrieval in databases**.

KEYWORDS

Text-based person search, graph model, adversarial learning

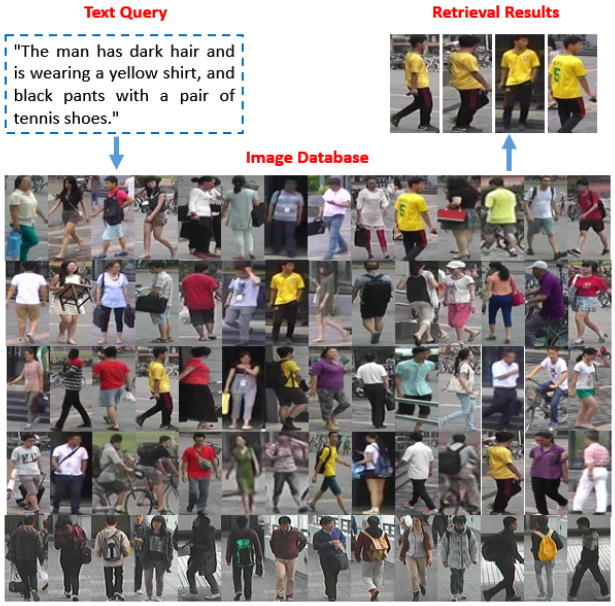

Figure 1: The illustration of text-based person search. Given a textual description of a pedestrian, the model aims to retrieve the corresponding pedestrian images from the image database.

ACM Reference Format:
Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. 2019. Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search. In *Proceedings of the 27th ACM Int'l Conf. on Multimedia (MM'19), Oct. 21–25, 2019, Nice, France.* ACM, Nice, France, 9 pages. https://doi.org/10.1145/3343031.3350991

1 INTRODUCTION

Person search aims at identifying a target pedestrian at diverse locations over different non-overlapping camera views by a given query. It has attracted increasing attention recently because of its importance for many practical applications, such as automated surveillance, activity analysis and criminal investigation [17, 30]. Existing methods of person search are mainly classified into three

*Corresponding author

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350991

categories according to the query type, *i.e.*, image-based query [21, 32, 36, 49], video-based query [3, 4, 22, 25, 39] and text-based query [5, 19, 47]. Image-based or video-based person search (person re-identification) requires at least one image or video of the target pedestrian, which in many cases is very difficult to obtain. Since textual descriptions of a pedestrian's appearance are more accessible, text-based person search can handle the problem of lacking person image/video.

The challenge of this task is to not only learn discriminative visual representation against background clutter, occlusion, the dramatic variations in illumination, body pose and camera viewpoint *etc* for pedestrians, but also effectively extract modality-invariant visual-textual representations in a shared feature space from raw images and textual descriptions of pedestrians [16]. Figure 1 illustrates one example of retrieval result of the text-based person search.

A few approaches have been proposed for text-based person search. Some of them [2, 5, 18, 19] attempt to build similarity models with the attention mechanism to compute the matching score of image-text pair. They utilize the attention mechanism to adaptively focus on the corresponding local regions of images and the words (or phrases) in the textual description under the same semantic relevances for learning fine-granularity features and fine matching. Nevertheless, their methods treat local patches of the images in isolation and only consider word-level context relations in the textual description. Thus, they neglect the structured semantic information contained in images and texts [6, 12, 13], which are able to provide more informative cues for distinguishing different pedestrians. Other methods mainly focus on learning joint embeddings from images and text in a latent common feature space [11, 47]. Two-branch architectures are exploited by them, where one branch extracts visual feature and another one extracts textual feature, and then the discriminative cross-modality embeddings are learned by the elaborate designed objective functions. The most common object functions for image-text matching include bi-directional ranking loss [24, 35] and canonical correlation analysis (CCA) [7, 41]. However, these object functions suffer from slow convergence and instability. In addition, these methods only consider paired image-text information, while neglect the overall data distribution within each modality which can be used to bridge the modality gap, result in unsatisfactory performance.

In this work, we propose a novel deep adversarial graph attention convolution network (A-GANet) to learn modality-invariant discriminative image-text representation for text-based person search. The A-GANet exploits directed semantic scene graph to generate fine-grained structured semantic representations from images and texts for matching, which includes the knowledge of present objects together with their attributes and spatial information, and pairwise relationships. It simultaneously utilizes adversarial learning to perform a min-max game between two process, *i.e.*, feature transformer and modality discriminator, for seeking an effective common feature space for the learned representations. As illustrated in Figure 2, A-GANet consists of an image graph attention network learning structured semantic visual feature, a text graph attention network extracting structured semantic textual feature and an adversarial learning module for learning an effective shared feature space. Specifically, the image graph attention network contains five residual blocks, a visual scene graph module and a joint embedding layer. The residual blocks are utilized to extract low-level visual feature. The visual scene graph module constructs a visual scene graph for learning the high-level structured visual feature with a novel graph attention convolution layer. The joint embedding layer is used to transfer the visual representation into an effective common feature subspace. The text graph attention network contains a bidirectional long short-term memory (Bi-LSTM), a textual scene graph module and a joint embedding layer. The Bi-LSTM is used to extract textual context feature. Moreover, the feature transformer in the adversarial learning module aims to generate modality-invariant representations to confuse the modality discriminator, while the modality discriminator acting as an adversary, attempts to distinguish the learned features from different modalities and indirectly steers the learning process of the feature transformer. By the joint adversarial learning on the two process, A-GANet is able to learn effective modality-invariant discriminative image-text representation, leading to satisfactory person retrieve results. We conduct extensive experiments to evaluate A-GANet on the two challenging benchmarks, and report superior performance over state-of-the-art approaches.

The main contribution of this paper is three-fold: (1) We propose a novel deep adversarial graph attention convolution network (A-GANet) for text-based person search. (2) We exploit semantic scene graphs to generate fine-grained structured representations from images and texts of pedestrians. (3) We bring the adversarial learning into the text-based person search for seeking an effective common feature space on the learned image and text features.

2 RELATED WORK

Recent years have witnessed many research efforts and encouraging progress on person search. This section briefly reviews existing works belonging to two major categories, *i.e.*, the newly emerging text-based person search methods and image/video based person search methods.

Text-based Person Search. Text-based person search retrieves the pedestrian through natural language description. Natural language can also depict a pedestrian as compared to images and videos with fewer restrictions. It is a good complement to image/video-based person search when corresponding image/video is difficult to obtain. Li *et al.* [19] proposed a GNA-RNN model to learn affinities between sentences and person images with a designed gated neural attention mechanism for text based person search. Zheng *et al.* [48] proposed a instance loss for image-text retrieval, which was based on an unsupervised assumption that every image/test group could be viewed as one class. Chen *et al.* [5] proposed an patch-word matching model, which computed the affinity between an image and a word and could accurately capture the local matching details between image and text. Zhang *et al.* [47] proposed a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss for learning discriminative image-text embeddings. The CMPM loss minimized the KL divergence between the projection compatibility distributions and the normalized matching distributions. The CMPC loss categorized the vector projection of features from one modality onto another

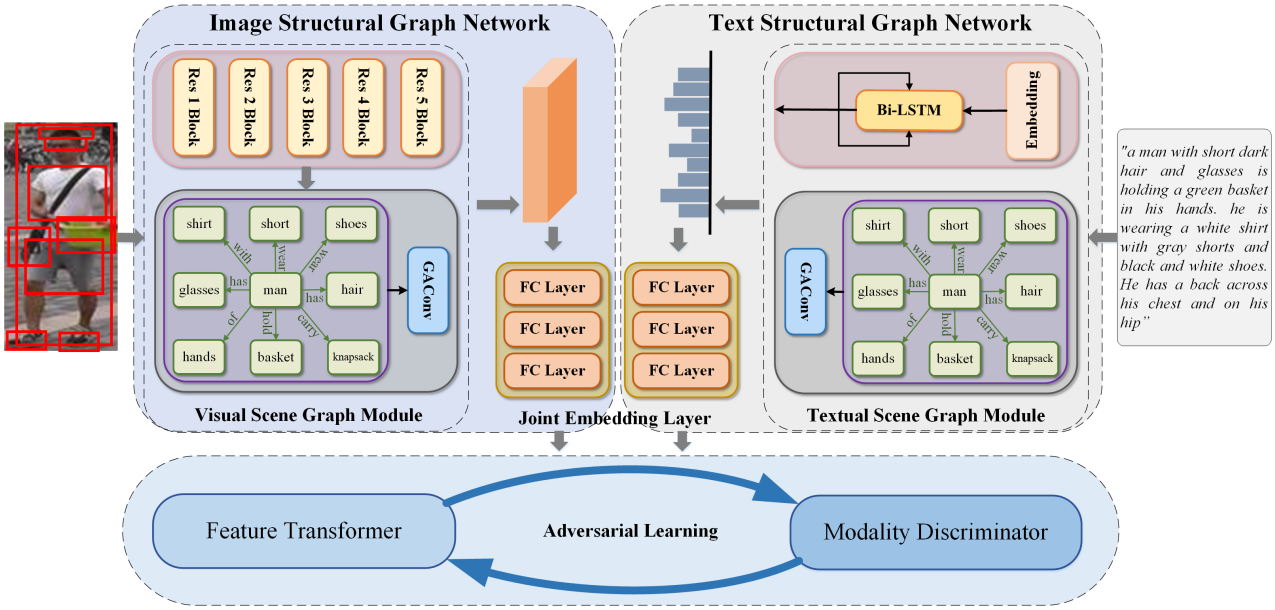

Figure 2: The overall architecture of the proposed A-GANet approach. It consists of an image graph attention network for extracting visual feature, a text graph attention network for learning textual feature as well as an adversarial learning module for seeking a common feature space.

with a improved norm-softmax loss. Li *et al.* [18] proposed an identity-aware two-stage framework for text-based person search, the stage-1 CNN-LSTM network learned to embed cross-modal features and provided initial training point for the stage-2 net with a Cross-Modal Cross-Entropy (CMCE) loss, the stage-2 CNN-LSTM network refined the matching results with a latent co-attention mechanism. Yamaguchi *et al.* [40] presented a model that combined methods for spatio-temporal person detection and multi-modal retrieval.

Image/video-based Person Search. For image/video-based person re-identification, deep learning based methods have shown substantial advantage over traditional hard-crafted features based methods or metric learning based methods on most of person search dataset. For example, Xiao *et al.* [37] presented a pipeline for learning global full-body representations from multiple domains by a Domain Guided Dropout layer to discard useless neurons for each domain. Liu *et al.* [23] proposed a multi-scale triplet CNN which captures visual appearance of a person at various scales by a comparative similarity loss on massive sample triplets. McLaughlin *et al.* [27] presented a recurrent neural network architecture for video-based person re-identification, which utilizes optical flow, recurrent layers and mean-pooling layer to learn video features containing appearance and motion information. Li *et al.* [20] formulated a method of jointly learning local and global features in a CNN model by optimizing multiple classification losses in different context. Shen *et al.* [31] proposed a Similarity-Guided Graph Neural Neural to incorporate the rich gallery-gallery similarity information into training process of person re-identification. Suh *et al.* [33] proposed a model for person re-identification, which consists of a two-stream network generating appearance and body part feature maps respectively, and a bilinear-pooling layer that fuses two feature maps to an image descriptor.

3 METHOD

In this section, we firstly present the overall architecture of the proposed approach, and then introduce each component of the architecture in the following subsections.

3.1 Overall Architecture

Given a training set $\{\boldsymbol{x}_i, \boldsymbol{t}_i\}_{i=1}^{N}$ containing N image samples $\boldsymbol{x}$ captured by non-overlapping camera networks and their corresponding textual descriptions $\boldsymbol{t}$ from K pedestrians together with their corresponding person ID as $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{N}$, the objective is to learn a modality-invariant discriminative representation for identifying the same pedestrian and distinguishing different pedestrians from the image gallery by using the query of textual description of a pedestrian. We propose a novel deep adversarial graph attention convolution network (A-GANet) for text-based person search, which exploits the directed semantic scene graph to extract fine-grained structured pedestrian representations from images and texts and bridge the modality gap, and simultaneously utilizes adversarial learning to project the two representations into a common latent feature space used for retrieving. As shown in Figure 2, A-GANet consists of an image graph attention network learning structured semantic visual feature, a text graph attention network extracting structure semantic textual feature and an adversarial learning module for learning an effective shared feature space. Specifically, the image graph attention network contains five residual blocks, a visual scene graph module and a joint embedding layer. The residual

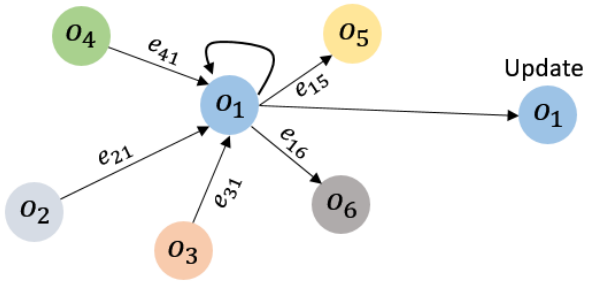

Figure 3: Schematic illustration of the proposed graph attention convolution layer.

blocks have strong ability in extracting visual representation. The visual scene graph module constructs a visual scene graph based on the object regions in the images and their relationships, and learn the high-level structured semantic visual feature from the graph with a designed graph attention convolution layer. The joint embedding layer containing three fully connected (FC) layers, which are used to transfer the representations into the effective common feature subspace. The text graph attention network contains a Bi-LSTM layer, a textual scene graph module and a joint embedding layer. The Bi-LSTM cell progressively takes the embedding of each word in the texts as input and captures both the historical and future contextual information of the processed words. Furthermore, the adversarial learning module contains a feature transformer and a modality discriminator, which is used to play a min-max game and optimize the extracted visual and textual features. The feature transformer is composed of a pairwise loss and an identification loss for effective feature learning. The modality discriminator contains an adversarial loss for distinguishing images and texts. During testing, the final similar score between the textual description t_i and the image x_i can be computed by the extracted textual representation $f(\boldsymbol{t}_i)$ and image representation $f(\boldsymbol{x}_i)$:

$$S(\boldsymbol{t}_i, \boldsymbol{x}_i) = \|f(\boldsymbol{t}_i) - f(\boldsymbol{x}_i)\|_2^2 \tag{1}$$

3.2 Image Graph Attention Network

The image graph attention network is based on ResNet-50 [9], consisting of five residual blocks, a visual scene graph module and a joint embedding layer, which is utilized to extract the visual feature from images. Since, the scene graph contains the structured semantic information of an image, the network with the visual scene graph module is able to capture more fine-grained visual cues and richer semantic understanding for the pedestrian, which can help shrink the modality gap between images and texts.

The images firstly go through the five residual blocks to obtain the initial feature map V_m. Each residual block contains a stack of 3 layers, The three layers are 1×1, 3×3, and 1×1 convolutions, where the 1×1 convolution layers are responsible for reducing and then increasing (restoring) feature size, leaving the 3×3 convolution layer with smaller input/output dimensions. Moreover, each convolution layer is followed by a Batch Normalization (BN) layer, Rectified Linear Units (ReLU) layer and optional Max-Pooling layer. After the five residual blocks, an average pooling operation is applied to the initial feature map for extracting the global appearance visual feature with 2048 dimension.

The visual scene graph $\mathcal{G}_v$ is built by the visual scene graph module. It is a tuple $\mathcal{G}_v = (O_v, \mathcal{E}_v)$ over object regions in the pedestrian images, where $O_v = \{o_i\}_{i=1}^N$ is the node set and $\mathcal{E}_v = \{e_{ij}\}$ is the directed edge set, respectively. The node o_i corresponds to the i-th object (or salient region), the edge e_{ij} denotes the relationship between the objects o_i and o_j, which is represented as triplets $< subject, predicate, object >$. Inspired by recent works in scene graph generation [38, 43, 45], we utilize Faster-RCNN [29] in conjunction with ResNet-50 model [9] as the object detector to detect objects in an image, Motifs relationship classifier [43] as the relationship classifier to predict the relationship between objects. In addition, we also use a designed attribute classifier to predict the attributes of the obtained objects [44, 46]. We extract four types of features, *i.e.*, appearance feature $\boldsymbol{v}_a$, spatial feature $\boldsymbol{v}_s$, label embedding $\boldsymbol{v}_l$ and attribute embedding $\boldsymbol{v}_{att}$ to form the node feature $\boldsymbol{v}_{o_i} = [\boldsymbol{v}_{a_i}; \boldsymbol{v}_{s_i}; \boldsymbol{v}_{l_i}; \boldsymbol{v}_{att_i}]$. The appearance feature is the extracted ROI [29] feature from the object region in the initial feature map $\boldsymbol{V}_m$, which is a 1024-dimensional vector. The spatial feature is 100-dimensional vector, which is generated from the encoded spatial information, *i.e.* top-left, bottom-right coordinates and the size of the object bounding box $[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{h\cdot w}{H\cdot W}]$ with respect to the whole image by a FC layer. The label embedding and attribute embedding are also 100-dimensional vectors. The edge feature $\boldsymbol{v}_{e_{ij}}$ is similar to $\boldsymbol{v}_{o_i}$, containing three types of features *i.e.*, appearance feature, spatial feature and label embedding, which are extracted from the union of the corresponding subject region and object region.

In order to extract the structured semantic visual feature from the directed graph, we develop a graph attention convolution (GAConv) layer as illustrated in Figure 3, which can operate natively on graphs. The traditional convolution layer takes a spatial grid of feature vectors as input, and produces a new spatial grid of vectors, where each output vector is a function of the local neighborhoods of its corresponding input vector. Thus, a convolution aggregates information across local neighborhood vectors of the input. Our graph attention convolution layer performs a similar function: given a directed graph with feature vector $\boldsymbol{v}_o$ of each object node, it computes new vectors $\boldsymbol{v'}_o$ for each object node. Besides, considering that other nodes have different effects on the target node, an attention mechanism is applied to the GAConv layer for computing the importance scores between object nodes. In the graph $\mathcal{G}_v$, o_i can denote "subject" or "object" in relationship triplets, which indicates that o_i will act as different roles due to different edge directions. Therefore, different functions should be designed to incorporate such knowledge. v_{o_i} can be updated as follows:

$$\begin{aligned}\boldsymbol{v}_{\boldsymbol{o_i}} = &\sum_{o_j \in sbj(o_i)} w_{ij} \cdot g_s(\boldsymbol{v}_{\boldsymbol{o_i}}, \boldsymbol{v}_{\boldsymbol{e_{ij}}}, \boldsymbol{v}_{\boldsymbol{o_j}}) \\ &+ \sum_{o_k \in obj(o_i)} w_{ik} \cdot g_o(\boldsymbol{v}_{\boldsymbol{o_k}}, \boldsymbol{v}_{\boldsymbol{e_{ki}}}, \boldsymbol{v}_{\boldsymbol{o_i}})\end{aligned} \tag{2}$$

where g_s and g_o denote the full-connected layers followed by a BN layer and ReLU layer. w_{ij} indicates the importance of the feature

of node j to node i, as well as w_{ik}. And the attention weight is computed as follows:

$$w_{i,j} = \frac{exp(\boldsymbol{w_a} \cdot \boldsymbol{v_{e_{ij}}} + b_a)}{\sum_j exp(\boldsymbol{w_a} \cdot \boldsymbol{v_{e_{ij}}} + b_a)} \tag{3}$$

To integrate the enhanced object node features, we add a virtual node to gather the features of all other object nodes by the graph attention operation as follows:

$$\boldsymbol{v_{o_v}} = \sum_{o_i \in O} w_i \cdot g_v(\boldsymbol{v_{o_i}}) \tag{4}$$

where g_v refers to a full-connected layers followed by a BN layer and ReLU layer. Attention score w_i are not calculated from edge feature. Instead, it is obtained by a linear transformation of the node feature:

$$w_i = \frac{exp(\boldsymbol{w_c} \cdot \boldsymbol{v_{o_i}} + b_c)}{\sum_i exp(\boldsymbol{w_c} \cdot \boldsymbol{v_{o_i}} + b_c)} \tag{5}$$

The virtual node feature $\boldsymbol{v_{o_v}}$ is the output of the visual scene graph module containing informative structured semantic knowledge. We contact the virtual node feature and the global appearance visual feature, and take them into the joint embedding layer, which consists of three 1024-dimensional full-connected layers for mapping the learned image feature of the pedestrian into the common feature space with respect to the pedestrian text feature. Figure 4 shows one example of the built visual semantic scene graph for a pedestrian image.

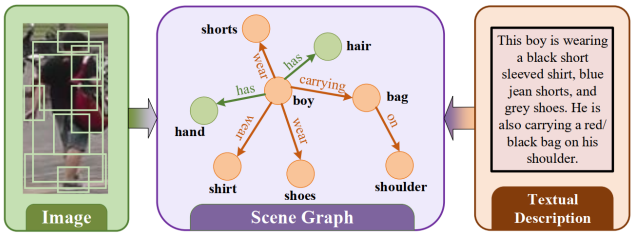

Figure 4: Visualization of semantic scene graph for a pair of image and text. The green and orange rounds and arrows are the nodes and edges of the built visual scene graph from the image. The visual scene graph is pruned to avoid clutter. The orange rounds and arrows are the nodes and edges of the parsed textual scene graph from the text.

3.3 Text Graph Attention Network

The text graph attention network is proposed to extract the effective textual feature from textual descriptions of pedestrians. It consist a Bi-LSTM layer, a textual scene graph module and a joint embedding layer.

Given a textual description, we firstly apply basic tokenizing and split it into words. Each word in the text is represented as an one-hot vector, then the one-hot vector is projected to a word embedding and sequentially processed with a Bi-LSTM, which is a combination of a forward LSTM and a backward LSTM. The Bi-LSTM memorizes the latent semantic dependencies among words, selectively discovers and propagate relevant historical and future context to next word. The hidden states of forward LSTM and backward LSTM are concatenated as the initial text representations with a max-pooling operation. In addition, the textual graph $\mathcal{G}_t$ is built by the textual scene graph module. We adopt the approach [1] to parse the textual scene graph. Analogously, it is a tuple $\mathcal{G}_t = (\mathcal{O}_t, \mathcal{E}_t)$ for describing objects and relationship between objects in the text, where $\mathcal{O}_t$ is the node set and $\mathcal{E}_t$ is the directed edge set, respectively. Different from the graph $\mathcal{G}_v$, the features of nodes in the $\mathcal{G}_t$ only include label embeddings and attribute embeddings, as well as the edge features contain label embeddings. We also employ a GAConv layer to update the nodes and obtain the structured semantic textual feature. After that, the structured semantic textual feature and the initial text feature are fused, and brought into the joint embedding layer for projecting the learned text feature of the pedestrian into the effective shared feature space with respect to the image feature. Figure 4 shows one example of the parsed textual semantic scene graph of a text.

3.4 Adversarial Learning Module

In order to learn more effective visual feature $\boldsymbol{v}$ and textual feature $\boldsymbol{t}$ in images and texts, the distributions of the two types of features should be modality-invariant and semantically discriminative, as well as maintain the underlying cross-domain semantic similarity between corresponding images and texts. Thus, we develop an adversarial learning module to optimize the extracted representations from the image graph attention network and the text graph attention network. It consists of a modality discriminator and a feature transformer, and plays a min-max game between the two process, similar to GAN [8], which can effectively and efficiently meet the above requirements for representations.

To close the modality gap between images and texts, the modality discriminator with parameters θ_D is introduced, which plays the role of "discriminator" in GAN. It contains three fully connected layers with 512, 256, 2-dimension, respectively. The goal is to distinguish the modality of a sample as reliably as possible given the trained feature transformer. An adversarial loss is used to classify the modality label (belonging to image or text) of the input sample, which is defined as follows:

$$\mathcal{L}_{adv}(\theta_D) = -\frac{1}{N}\sum_{i=1}^{N}(\log D(v_i;\theta_D) + log(1 - D(t_i;\theta_D))) \tag{6}$$

where $D(\cdot;\theta_D)$ refers to the calculated modality probability of the input sample i by the three fully connected layers.

In addition, the feature transformer is proposed to transfer the modality-invariant features of texts and images into a common subspace, which is composed of an identification loss and a pairwise loss. It acts as the role of "generator" in GAN. The identification loss enables the image representations and text representations to be discriminative in the common feature space, given person IDs. The pairwise loss ensures that the representations belonging to the same person ID possess high semantic similarity and invariant across modalities, and are projected into the shared feature space. The identification loss is formulated as follows:

$$\mathcal{L}_{ide}(\theta_V, \theta_T) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\boldsymbol{W}_{y_i}^T \boldsymbol{x}_i + \boldsymbol{b}}}{\sum_{j=1}^{K} e^{\boldsymbol{W}_j^T \boldsymbol{x}_j + \boldsymbol{b}}} \tag{7}$$

where y_i is the corresponding person ID of the i-th sample (image or text), $\boldsymbol{x_i}$ refers to image feature $\boldsymbol{v}$ or text feature $\boldsymbol{t}$, θ_V and θ_T are the parameters of the proposed image and text graph attention networks. $\boldsymbol{W}_j \in \mathbb{R}^{1024}$ represents the j-th column of the weight matrix $\boldsymbol{W} \in \mathbb{R}^{1024\times 11003}$ and $\boldsymbol{b}$ refers to a bias term. The pairwise loss is formulated as follows:

$$\mathcal{L}_{pair}(\theta_V, \theta_T) = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{\boldsymbol{W}_{p,y_i}^T \boldsymbol{z}_i + \boldsymbol{b}_p}}{\sum_{j=1}^{2} e^{\boldsymbol{W}_{p,j}^T \boldsymbol{z}_j + \boldsymbol{b}_p}} \tag{8}$$

where y_i is a 2-dim vector, indicating that the input pairs of images and texts whether have the same person ID. $\boldsymbol{z_i}$ refers to the fused feature by concatenating the image feature $\boldsymbol{v}$ and the text feature $\boldsymbol{t}$, $\boldsymbol{W}_{p,j}$ represents the j-th column of the weight matrix $\boldsymbol{W}_p$. Thus, the total loss for the feature transformer is $\mathcal{L}_{fea}(\theta_V, \theta_T) = \mathcal{L}_{pair} + \alpha \cdot \mathcal{L}_{ide}$, referred to feature projecting loss.

The process of optimizing the image and text representations is conducted by jointly optimizing the adversarial loss and the feature projecting loss. Nevertheless, the optimization goals of the two objective functions are opposite, the process runs a min-imax game for the feature transformer and the modality discriminator:

$$\begin{aligned}(\hat{\theta}_V, \hat{\theta}_T) &= \arg\min_{\theta_V, \theta_T} \ (\mathcal{L}_{fea}(\theta_V, \theta_T) - \mathcal{L}_{adv}(\theta_D)) \\ \hat{\theta}_D &= \arg\max_{\theta_D} \ (\mathcal{L}_{fea}(\theta_V, \theta_T) - \mathcal{L}_{adv}(\theta_D))\end{aligned} \tag{9}$$

The whole optimization process of the above two loss functions for parameters θ_V, θ_T and θ_D is similar to GAN.

4 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of the proposed A-GANet on the challenging text-based person search dataset *i.e.*, CUHK-PEDES and compare the A-GANet to state-of-the-art methods. To make our method more convincing and generic, we also conduct experiments on another cross-modality retrieval dataset *i.e.*, Flickr30K. The experimental results show that A-GANet achieves superior performance of retrieval over the state-of-the-art methods. Moreover, we investigate the effectiveness of each component of A-GANet, including the image graph attention network, the text graph attention network and the adversarial learning module.

Datasets- CUHK-PEDES [19] is the only one dataset for text-based person search, currently. It contains 40,206 pedestrian images of 13,003 identities. Each pedestrian image is described by two different textual descriptions. The dataset is splitted into three subsets for training, validation, and test, without having overlaps with same person IDs. The training set contains 11,003 pedestrians, 34,054 images and 68,108 sentence descriptions. The validation set and test set contain 3,078 and 3,074 images, respectively, both of which have 1,000 pedestrians. Flickr30k [42] is one of the largest cross-modality retrieval datasets. It contains 31,783 images collected from Flickr website, where each image has five text descriptions.

Table 1: Performance comparison to the state-of-the-art methods on the CUHK-PEDES dataset.

Method	Text-to-Image		
	Rank-1	Rank-5	Rank-10
CNN-RNN [28]	8.07	-	32.47
Neural Talk [34]	13.66	-	41.72
GNA-RNN [19]	19.05	-	53.64
IATVM [18]	25.94	-	60.48
PWM-ATH [5]	27.14	49.45	61.02
GLA [2]	43.58	66.93	76.26
CAN [11]	45.52	67.12	76.98
CMPM-CMPC [47]	49.37	-	79.27
A-GANet	**53.14**	**74.03**	**81.95**

We follow the protocol in [47] to split the dataset into 29,783 images for training, 1,000 images for validation, 1,000 images for testing.

Implementation Details- The implementation of the proposed method is based on the Pytorch framework with four NVIDIA Titan XP GPUs. The adam optimization algorithm is started with learning rate lr of 0.0002, the weight decay of $5e^{-4}$, the parameter α in the feature projecting loss is set to 1. All the images are resized to the size of $384 \times 192 \times 3$ and normalised with $1.0/256$. Meanwhile, the training set is enlarged by data augmentation of random horizontal flipping during training phase. The number of mini-batches is set to 100. The proposed network is optimized for 50 epochs in total. For visual scene graph module, we pre-train the object detector and Motifs relationship classier, as well as the attribute classier which is an fc(1024-d)-ReLU-fc(512-d)-Softmax network, on the Visual Genome Dataset [15]. This dataset has abundant scene graph annotations, *e.g.*, categories and attributes of objects, and the relationships between objects. Since the annotations of object, attribute, and relationship in the dataset are very noisy, we filter them by deleting the objects, attributes, and relationships which appear less than 2,000 times, and manually select 200 objects, 100 attributes and 50 relationships used to train the above three classifiers, both of which are also most frequently appeared in the CUHK-PEDES dataset, simultaneously. The pre-trained three classifiers are utilized to build the visual scene graph for each pedestrian image in CUHK-PEDES dataset. The word embedding matrix used in the visual scene graph module is $\boldsymbol{W}_{e_v} \in \mathbb{R}^{100\times 350}$. The dimensions of the FC layers used in Eq (2), (4) are 1024. For textual scene graph module, we use the method [1] to parse the textual scene graph for the texts in the CUHK-PEDES dataset, and filter the words which appears less than 3 times. After filtering, 4000 unique words are obtained. The word embedding matrix for textual scene module is $\boldsymbol{W}_{e_t} \in \mathbb{R}^{512\times 4000}$. The size of Bi-LSTM is set to 512, and the dimensions of the FC layers used in textual scene graph module are set to 512. The final extracted image and text features are 1024-dimension vectors.

Protocol- We adopt the top-k accuracy to evaluate the performance of person search. Given a query of textual description, all test images are ranked according to their similarity scores with the query. if any pedestrian image with the same person ID are contained in the top-k images, a successful search is achieved. Top-k

Table 2: Performance comparison to the state-of-the-art methods on the Flickr30k dataset.

Method	Text-to-Image		
	Rank-1	Rank-5	Rank-10
DCCA [41]	12.60	31.00	43.00
DVSA [14]	15.20	37.70	50.50
m-CNN [26]	26.20	56.30	69.60
DSPE [35]	29.70	60.10	72.10
sm-LSTM [10]	30.20	60.04	72.30
RRF-Net [24]	35.40	68.30	79.90
CMPM-CMPC [47]	37.30	65.70	75.5
A-GANet	**39.52**	**69.91**	**80.91**

accuracy thus represents the percentage of successful retrieve for a specific k.

4.1 Comparison to State-of-the-Arts

CUHK-PEDES: Table 1 shows the performance comparison of the proposed A-GANet against 8 state-of-the-art methods in terms of top-k accuracy. The compared methods belong to two categories, *i.e.*, building similarity network with attention mechanism methods, including CNN-RNN [28] , Neural Talk [34] , GNA-RNN [19] , IATVM [18] , PWM-ATH [5] , GLA [2] , and focusing on joint embedding methods, including CAN [11] and CMPM-CMPC [47]. The proposed A-GANet achieves 53.14%, 74.03%, 81.95% of rank-1, rank-5, rank-10 accuracy, respectively. We can see that our method surpasses existing methods, improving the 2nd best compared method CMPM-CMPC by 3.77% rank-1 recognition rate, which demonstrates the effectiveness of the proposed method. Moreover, A-GANet achieves significant performance improvement as compared to the building similarity network with attention mechanism methods, which indicates that the proposed A-GANet is able to capture more effective fine-grained structured image and text information. An illustration of some retrieval results is given in Figure 6.

Flickr30k: For Flickr30k dataset, we compare our method with 7 state-of-the-art approaches, including DCCA [41], DVSA [14], m-CNN [26], DSPE [35], sm-LSTM [10], RRF-Net [24] and CMPM-CMPC [47]. As shown in Table 2, Our method obtains 39.52%, 69.9%, 80.9% of rank-1, rank-5, rank-10 recognition rate, respectively. It improves the second best result of CMPM-CMPC method by 2.22% rank-1 accuracy. Further, it can be observed that our method outperforms these methods at all ranks of recognition rate, which validates the effectiveness of our method not only for text-based person re-identification, but also for the cross-modal retrieval task.

4.2 Ablation Studies

To demonstrate the effectiveness and contribution of each component of the A-GANet, we conduct a series of ablation experiments on the CUHK-PEDES dataset. We evaluate the effect of the visual scene graph module, the textual scene graph module and the adversarial loss module, and compare their performance for text-based person search.

Table 3 summarizes the ablation results of the proposed A-GANet. A-GANet w/o vis refers to A-GANet without visual scene graph module, which only exploits the global appearance visual feature extracted from the five residual blocks as the final image feature for matching. A-GANet w/o tex refers to A-GANet without textual scene graph module, which takes the output of Bi-LSTM as the final textual feature. A-GANet w/o adv refers to A-GANet without adversarial loss, and only using the identification loss and the pairwise loss to optimize the model. A-GANet refers to the whole framework of the proposed method. From Table 3, A-GANet w/o vis obtains 48.26 % rank-1 accuracy, A-GANet w/o tex achieve 49.51 % rank-1 accuracy and A-GANet w/o adv achieve 50.65 % rank-1 accuracy, respectively. Moreover, A-GANet yields the best performance of 53.14% of rank-1 recognition rate. By comparing A-GANet w/o vis with A-GANet, we can observe that the visual scene graph module is able to capture more useful structured semantic visual cues and learn more effective image feature. By comparing A-GANet w/o tex with A-GANet, we can observe that the text scene graph module can absorb structured semantic textual information and learn more effective text feature. In addition, the A-GANet outperforms A-GANet w/o adv by 2.49% top-1 metric, which proves the effectiveness of adversarial loss module for learning an effective shared feature space.

Table 3: Evaluation of the effectiveness of each component within A-GANet on the CUHK-PEDES dataset.

Method	Text-to-Image		
	Rank-1	Rank-5	Rank-10
A-GANet w/o vis	48.26	72.01	80.65
A-GANet w/o tex	49.51	73.51	81.66
A-GANet w/o adv	50.65	73.33	81.11
A-GANet	53.14	74.03	81.95

Table 4: Evaluation of the effectiveness of the GAConv layer within the scene graph modules on the CUHK-PEDES dataset.

Method	Text-to-Image		
	Rank-1	Rank-5	Rank-10
A-GANet w/o visAtt	50.42	73.31	81.10
A-GANet w/o texAtt	51.43	73.82	81.47
A-GANet	53.14	74.03	81.95

Table 5: Evaluation of the effectiveness of each component within the adversarial learning module on the CUHK-PEDES dataset.

Method	Text-to-Image		
	Rank-1	Rank-5	Rank-10
A-GANet with $\mathcal{L}_{ide}$	28.88	56.91	69.65
A-GANet with $\mathcal{L}_{pair}$	41.04	64.44	74.76
A-GANet with $\mathcal{L}_{fea}$	50.65	73.33	81.11
A-GANet	53.14	74.03	81.95

Analysis of the scene graph module. Table 4 summarizes the ablation results of the attention mechanism of the GAConv layer

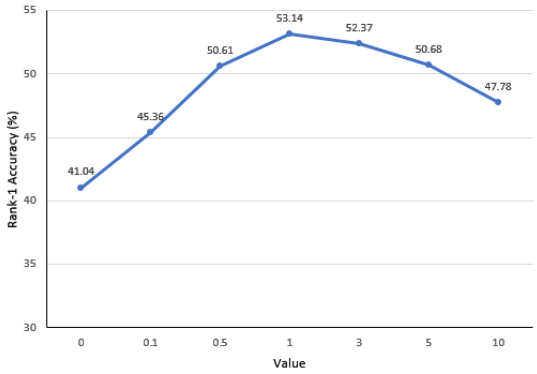

Figure 5: Evaluation of the proposed adversarial learning module with different values of parameter α.

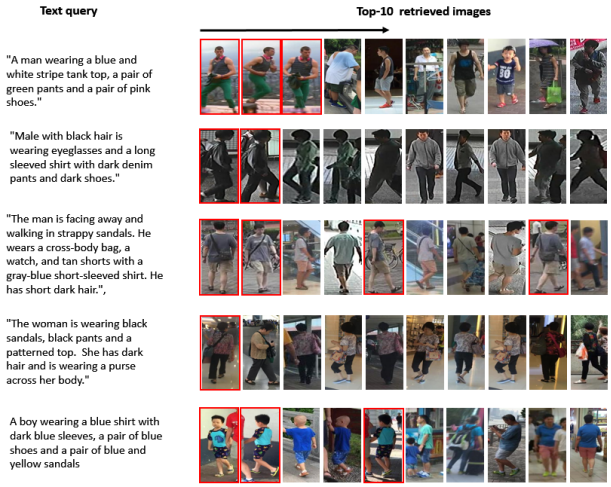

Figure 6: Examples of top-10 retrieved images based on the text query by the proposed A-GANet. Retrieval results are sorted by their similarity scores with the texts. Red boxes indicate the corresponding images for the texts.

in the proposed scene graph module. A-GANet w/o visAtt refers to A-GANet replacing the attention mechanism in the visual scene graph module with an average pooling operation. A-GANet w/o texAtt refers to A-GANet replacing the attention mechanism in the textual scene graph module with an average pooling operation. A-GANet refers to the whole framework of the proposed method. From table 4, A-GANet w/o visAtt and A-GANet w/o texAtt obtain 50.42 % and 51.43% rank-1 accuracy, respectively. The performance improvement of A-GANet over A-GANet w/o visAtt and A-GANet w/o texAtt by 2.72% and 1.71% rank-1 accuracy respectively, indicates that the attention mechanism in the scene graph modules can extract more precise visual and textual information when propagating messages between connecting nodes, and thus benefits the performance.

Analysis of the adversarial loss module. Table 5 compares each component of the adversarial loss module. A-GANet with $\mathcal{L}_{ide}$ refers to A-GANet only using the identification loss. A-GANet with $\mathcal{L}_{pair}$ refers to A-GANet only using the pairwise loss. A-GANet with $\mathcal{L}_{fea}$ refers to A-GANet using both the identification loss and the pairwise loss. A-GANet refers to the whole framework of the proposed method. From table 5, A-GANet with $\mathcal{L}_{ide}$ and A-GANet with $\mathcal{L}_{pair}$ obtain 28.88% and 41.04% rank-1 accuracy, respectively. We can obverse that the performance of A-GANet with $\mathcal{L}_{ide}$ is inferior to A-GANet with $\mathcal{L}_{pair}$, indicting that the pairwise loss can extract more effective features by minimizing the gap among the representations of the matching pairs of images and texts. Meanwhile, A-GANet with $\mathcal{L}_{fea}$ obtains better performance than both of A-GANet with $\mathcal{L}_{ide}$ and A-GANet with $\mathcal{L}_{pair}$. It indicates that both the intra-modal discriminativeness and inter-modal invariance terms contribute to the final retrieval accuracy. A-GANet yields the best performance of 53.14% of rank-1 recognition rate, which shows the effectiveness of A-GANet for joint optimizing of both the adversarial loss and the feature projecting loss. In addition, the parameters α are key parameter for the feature projecting loss, which controls the relative importance between pairwise loss and the identification loss. We conduct experiment to evaluate the impact of α, the results are shown in Figure 5. From Figure 5, we can see that when $\alpha = 1$, A-GANet yields the best retrieval performance.

5 CONCLUSIONS

In this work, we propose a novel deep adversarial graph attention convolution network (A-GANet) to learn discriminative and modality-invariant representations for text-based person search. The image and text graph attention networks utilize the scene graph modules to build scene graphs of images and text. The graph attention networks can dynamically focus on the more informative objects in the scene graphs and then learn discriminative structured semantic representations, including the information of present objects with their attributes and spatial position, as well as pairwise relationships, by the graph attention convolution layers. An adversarial learning module is designed to perform a min-max game between two process, *i.e.*, feature transformer and modality discriminator, to learn an effective common space for the image and text representations. We conducted extensive experiments on two widely-used datasets, *i.e.*, CUHK-PEDES and Flickr30k. The experimental results have shown that the proposed A-GANet improves the performance of person search over a wide range of state-of-the art methods.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants 61622211, 61620106009 and 61525206 as well as the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*. Springer, 382–398.
[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. Improving deep visual representation for person

re-identification by global and local image-language association. In *Proceedings of the European Conference on Computer Vision*. 54–70.
[3] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. 2018. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1169–1178.
[4] Di Chen, Zheng-Jun Zha, Jiawei Liu, Hongtao Xie, and Yongdong Zhang. 2018. Temporal-contextual attention network for video-based person re-identification. In *Proceedings of the Pacific Rim Conference on Multimedia*. Springer, 146–157.
[5] Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision*. 1879–1887.
[6] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. 2018. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7239–7248.
[7] Matthias Dorfer, Jan Schlüter, Andreu Vall, Filip Korzeniowski, and Gerhard Widmer. 2018. End-to-end cross-modality retrieval with CCA projections and pairwise ranking loss. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 117–128.
[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing System*. 2672–2680.
[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[10] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2310–2318.
[11] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Cascade Attention Network for Person Search: Both Image and Text-Image Similarity Selection. *arXiv preprint arXiv:1809.08440* (2018).
[12] Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1219–1228.
[13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3668–3678.
[14] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.
[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.
[16] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*. 201–216.
[17] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 369–378.
[18] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*. 1890–1899.
[19] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1970–1979.
[20] Wei Li, Xiatian Zhu, and Shaogang Gong. 2017. Person Re-Identification by Deep Joint Learning of Multi-Loss Classification. In *Proceeding of the International Joint Conference on Artificial Intelligence*. 2194–2200.
[21] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. 2019. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7202–7211.
[22] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. 2019. Dense 3D-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1s (2019), 8.
[23] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-scale triplet cnn for person re-identification. In *Proceedings of the ACM Conference on Multimedia Conference*. ACM, 192–196.
[24] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 4107–4116.
[25] Zimo Liu, Dong Wang, and Huchuan Lu. 2017. Stepwise metric promotion for unsupervised video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 2429–2438.
[26] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision*. 2623–2631.
[27] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. 2016. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1325–1334.
[28] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–58.
[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems*. 91–99.
[30] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 420–429.
[31] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. 2018. Person Re-identification with Deep Similarity-Guided Graph Neural Network. In *Proceedings of the European Conference on Computer Vision*. 486–504.
[32] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1179–1188.
[33] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision*. 402–419.
[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.
[35] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5005–5013.
[36] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. 2018. Person re-identification with cascaded pairwise convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1470–1478.
[37] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1249–1258.
[38] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5410–5419.
[39] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 4733–4742.
[40] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Spatio-Temporal Person Retrieval via Natural Language Queries. In *Proceedings of the IEEE International Conference on Computer Vision*. 1453–1462.
[41] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3441–3450.
[42] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.
[43] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5831–5840.
[44] Hanwang Zhang, Zheng-Jun Zha, Shuicheng Yan, Jingwen Bian, and Tat-Seng Chua. 2012. Attribute feedback. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 79–88.
[45] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, and Tat-Seng Chua. 2014. Robust (semi) nonnegative graph embedding. *IEEE Transactions on Image Processing* 23, 7 (2014), 2996–3012.
[46] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 33–42.
[47] Ying Zhang and Huchuan Lu. 2018. Deep Cross-Modal Projection Learning for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision*. 686–701.
[48] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding with Instance Loss. *arXiv preprint arXiv:1711.05535* (2017).
[49] Xierong Zhu, Jiawei Liu, Hongtao Xie, and Zheng-Jun Zha. 2019. Adaptive alignment network for person re-identification. In *Proceedings of the International Conference on Multimedia Modeling*. Springer, 16–27.