Adversarial Attribute-Text Embedding for Person Search with Natural Language Query

Zheng-Jun Zha *Member, IEEE,* Jiawei Liu, Di Chen and Feng Wu *Fellow, IEEE,*

***Abstract*—The newly emerging task of person search with natural language query aims at retrieving the target pedestrian by a text description of the pedestrian. It is more applicable compared to person search with image/video query, *i.e.*, person re-identification. In this paper, we propose a new Adversarial Attribute-Text Embedding (AATE) network for person search with text query. In particular, a cross-modal adversarial learning module is proposed to learn discriminative and modality-invariant visual-textual features. It consists of a cross-modal learner and a modality discriminator, playing a min-max game in an adversarial learning way. The former is to improve intra-modality discrimination and inter-modality invariance towards confusing the modality discriminator. The latter is to distinguish the features from different modalities and boost the learning of modality-invariant features. Moreover, an attribute graph convolutional network is proposed to learn visual attributes of pedestrians, which possess better descriptiveness, interpretability and robustness compared to pedestrian appearance features. A hierarchical text embedding network, consisting of multi-stacked bidirectional LSTMs and an attention block, is developed to extract effective textual features from text descriptions of pedestrians. Extensive experimental results on two challenging benchmarks, have demonstrated the effectiveness of the proposed approach.**

***Index Terms*—person search, natural language, adversarial learning, visual attributes, graph convolution network,**

I. Introduction

PERSON search aims to find a target pedestrian from images/videos taken by non-overlapping cameras, given a text or image/video query. It has attracted increasing attention recently due to its broad prospects in many practical applications, such as automated surveillance, activity analysis and criminal investigation *etc.* [1]–[6]. Existing person search methods can be mainly divided into three categories according to the type of query, *i.e.*, image-based [7]–[10], video-based [11]–[13] and text-based person search [14]–[16]. Image/video-based person search, termed as person re-identification in literature, requires at least one image or video clip of the target pedestrian as query, which is difficult to obtain in practice. In contrast, text description of a pedestrian is more accessible. Hence, text-based person search is more applicable as compared to image/video based search. However, it is more challenging due to that it requires not only discriminative visual and textual features of pedestrians, but also cross-modal matching [17]. Figure 1 shows an illustration of person search with text description.

Zheng-Jun Zha, Jiawei Liu, Di Chen and Feng Wu are with the School of Information Science and Technology, University of Science and Technology of China, Hefei, 230027, China. Corresponding author: Zheng-Jun Zha, Email: zhazj@ustc.edu.cn.

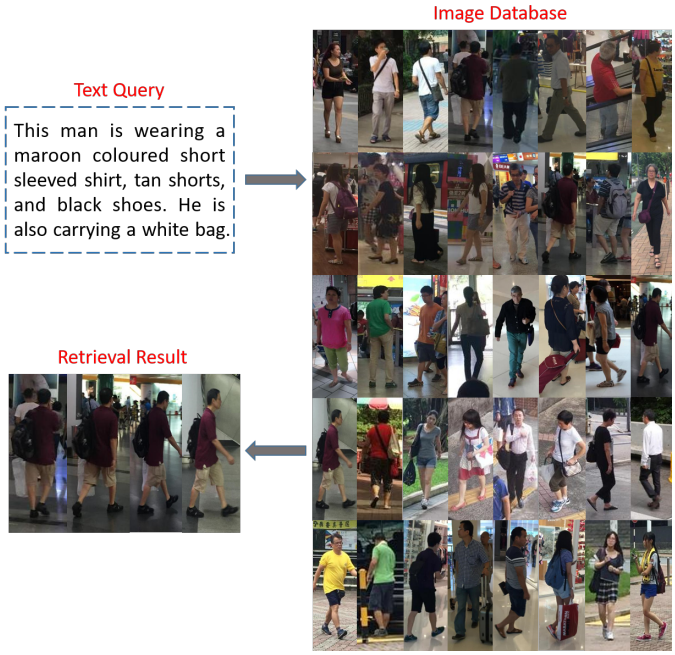

Fig. 1. The illustration of person search with natural language query. Given a text description of a pedestrian, it is to retrieve the relevant pedestrian images from gallery.

A few of works have been proposed for person search with natural language query in recent years and have steadily improved the performance [14]–[16], [18]–[22]. Most of them proposed to learn cross-modal affinity between image and text by the designed similarity learning networks [15], [16], [19]–[21]. They mainly learn the matching between image region and word/phrase through attention mechanism, and then match image and text description by aggregating region-word/phrase matching. Other works proposed to learn a latent common feature space of visual and textual modalities, towards boosting the cross-modal matching between pedestrian image and text description [14]. They exploit a two-branch network to extract visual and textual features respectively, and embed them into a common space by the designed loss functions. The widely used loss functions include identification loss [19], [21], bi-directional ranking loss [23], [24] and Canonical Correlation Analysis (CCA) loss [25], [26]. While existing works preserve semantic affinity in cross-modal similarity or common feature space, this work further pursues modality-invariance of visual and textual representation. That is, the distributions over visual and textual modalities are encouraged to be similar in the

learned representation. Moreover, existing methods are based on visual appearance feature of pedestrian, which is often not robust against background clutter, variations in illumination, body pose and viewpoint *etc*.

In this work, we propose a new adversarial attribute-text embedding (AATE) network for person search with natural language query. In particular, we propose a cross-modality adversarial learning module, consisting of a cross-modal learner and a modality discriminator. The perform a adversarial min-max game, towards learning effective visual and textual features. The learned features are not only discriminative within each modality but also of similar distributions across modalities. As a text description usually describes various attributes of the pedestrian, the learning of visual attributes is crucial for text-image matching. Visual attributes, as intermediate-level semantic descriptor, possess better robustness against the variations of illumination, body pose and camera viewpoint, compared to visual appearance of pedestrians. Hence, we propose a visual attribute graph convolutional network to learn visual attributes of pedestrians. We also develop a hierarchical text embedding network consists of multi-stacked bidirectional LSTMs and a textual attention block for learning effective textual features.

As illustrated in Figure 2, AATE consists of a visual attribute graph network, a hierarchical text embedding network and an adversarial learning module. In particular, the attribute graph network first computes latent spatial attention on each attribute and concentrates the network on corresponding image regions during learning attributes. It then exploits the semantic context among various attributes by a graph convolutional network to improve attribute learning. The hierarchical text embedding network contains multi-stacked bidirectional LSTMs (Bi-LSTMs) and a textual attention block. The former exploits the context among words, while the latter discovers informative words, towards learning precise and discriminative textual features. In the cross-modal adversarial learning module, the cross-modal learner improves intra-modality discrimination and inter-modality invariance towards confusing the modality discriminator, while the discriminator, acting as an adversary, attempts to distinguish the learned features from different modalities. We conduct extensive experiments to evaluate AATE on 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 new adversarial visual attribute and text embedding network for person search with natural language query; (2) we develop a cross-modal adversarial learning module to learn effective visual-textual representation, which is not only discriminative within each modality but also has similar distribution across modalities; (3) we use an attribute graph network, which learns effective attribute features by jointly exploiting visual attention on attributes as well as semantic dependencies among attributes.

II. Related Works

Recent years have witnessed many research efforts and encouraging progress on the task of person search. Existing works can be roughly divided into two categories, *i.e.*, the newly emerging text-based person search and the conventional image/video-based person search, *i.e.*, person re-identification. Text-based person search aims to retrieve the target pedestrian through a natural language description. It is also related to the conventional cross-modal retrieval [27]–[29] . Here, we mainly review the works of person search.

Text-based Person Search. Most of existing works focus on learning cross-modal affinity between image and text by the designed similarity learning networks [14]–[16], [19]–[22]. For example, Li *et al.* [15] proposed to learn affinities between text sentences and pedestrian images with a designed gated neural attention mechanism. Chen *et al.* [16] proposed a patch-word matching model, which computes the affinity between an image patch and a word. It captures local matching details between image and text. Li *et al.* [19] proposed an identity-aware two-stage CNN-LSTM framework for text-based person search. The first CNN-LSTM is to embed multi-modal features with a cross-modal cross-entropy loss. The second CNN-LSTM refines the image-text matching with a latent co-attention mechanism. Yamaguchi *et al.* [30] presented a person search method that combines the modes for spatio-temporal person detection and multi-modal retrieval. Liu *et al.* [22] proposed a deep adversarial graph attention convolution network (A-GANet) for text-based person search. It exploited textual and visual scene graphs to improve the descriptiveness of textual and visual features. The scene graphs [31]–[33] model the relation among words in text description as well as the dependency among bounding boxes detected within images, respectively. Some other methods proposed to learn a latent common feature space of visual and textual modalities. For example, Zhang *et al.* [14] proposed a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss for learning discriminative image-text embedding. The CMPM loss minimizes the KL divergence between the projection compatibility distributions and the normalized matching distributions. The CMPC loss categorizes the vector projection of features from one modality to another with a norm-softmax loss. Zheng *et al.* [18] proposed an instance loss for cross image-text retrieval based on an assumption that every image/text group could be viewed as one class.

Image-based Person Search. Image-based person search matches a query pedestrian image by one camera with gallery images from other non-overlapping cameras. Conventional methods mainly focus on either extracting discriminative features [34]–[38] or learning appropriate distance metrics [39]–[43]. Recently, with the great progress of deep neural network, deep learning based methods have shown substantial advantage over traditional hard-crafted features or metric learning based methods on most of the person re-identification benchmarks. For example, Ahmed *et al.* [44] proposed a specifically designed CNN and a binary verification loss which takes a pair of cropped pedestrian images as input for person re-identification. Ding *et al.* [45] introduced a triplet loss to minimize feature distance between the same pedestrian and maximize the distance among different pedestrians during training. Liu *et al.* [46] proposed a multi-scale triplet CNN

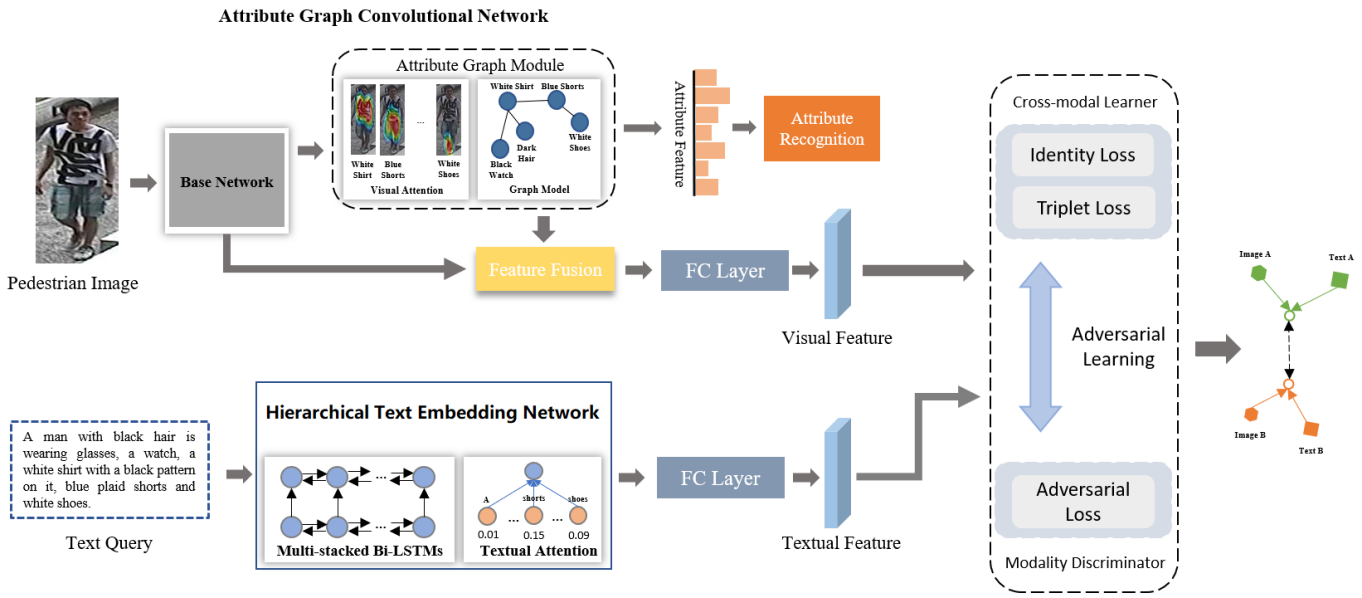

Fig. 2. The overall architecture of the proposed AATE network. It consists of a visual attribute graph convolutional network for learning visual feature, a hierarchical text embedding network for learning textual feature as well as a cross-modal adversarial learning module for learning modality-invariant and discriminative visual-textual representation.

which captures visual appearance of a person at multiple scales by a comparative similarity loss on sample triplets.

Video-based Person Search. Video-based person search is an extension of image-based person search. It searches for the target pedestrian by one or multiple video clips of the pedestrian [47]–[51]. Compared to images, video sequences contain motion patterns of pedestrians as well as more presence of pedestrian appearance. Early methods are based on hand-crafted video representations and/or appropriate distance metric. For example, You *et al.* [49] developed a top-push distance learning model to optimize the matching accuracy of top rank results. Recent works proposed deep learning models for video-based person search [47], [52]–[55]. For example, Liu *et al.* [52] proposed a Dense 3D-Convolutional Network (D3DNet), which introduces multiple 3D dense blocks to learn spatio-temporal and appearance features of pedestrians. McLaughlin *et al.* [47] presented a recurrent neural network architecture, which consists of optical flow, recurrent layers and mean-pooling layer to learn visual appearance and motion features of pedestrians. Li *et al.* [53] proposed to jointly learn local and global features in a CNN model by optimizing multiple classification losses in different context. Shen *et al.* [54] proposed a similarity-guided graph neural network to incorporate gallery-gallery similarities into the training process of person re-identification model.

III. Method

Supposing a training set $\{I_i, T_i\}_{i=1}^N$, it includes N pairs of pedestrian image and text description, where $\{I_i\}_{i=1}^N$ are pedestrian images taken by non-overlapping cameras and $\{T_i\}_{i=1}^N$ are the corresponding text descriptions of pedestrians. $Y = \{y_i\}_{i=1}^N$, where $y_i \in [1, 2, \cdots, K]$ is pedestrian ID. The task is to identify the target pedestrian images in gallery based on a text query. Figure 2 illustrates the architecture of the proposed adversarial attribute-text embedding (AATE) network, consisting of a visual attribute graph neural network, a hierarchical text embedding network and a cross-modal adversarial learning module.

A. Visual Attribute Graph Convolutional Network

Text description usually describes multiple attributes of the target pedestrian. Hence, detecting visual attributes of pedestrian is of great importance for searching pedestrian. Moreover, visual attributes possess better descriptiveness, interoperability and robustness as compared to appearance feature. An attribute usually arises from one or more regions rather than the entire pedestrian image. It is thus necessary to concentrate on related regions during attribute learning. Moreover, different attributes correlate semantically. The presence or absence of a certain attribute is usually useful for inferring the presence/absence of other related attributes. For example, "wearing a dress" and "long hair" are likely to co-occur, while "carrying a bag" and "carrying a backpack" may mutually exclusive. Based on the above observation, we develop a visual attribute graph convolutional network to learn effective attribute features of pedestrians. As illustrated in Figure 3, the network consists of a visual attention block and a graph convolutional network. The visual attention block infers spatial attention for each attribute and concentrates the network on the corresponding local regions during attribute learning. The graph network exploits the underlying semantic dependencies among attributes which could effectively boost the learning of attributes [56]–[58].

The ResNet-50 [59] is used as the base network to extract feature map $\mathbf{V}_r$ from input image. The dimension of $\mathbf{V}_r$ is $7 \times 7 \times 2048$. The appearance feature with 2,048 dimension is

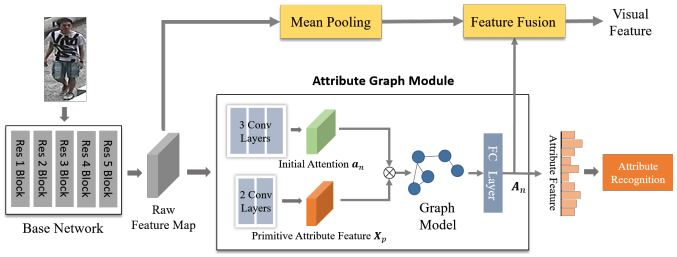

Fig. 3. Detailed structure of the visual attribute graph convolutional network.

generated after a mean pooling on $\mathbf{V}_r$. Taking $\mathbf{V}_r$ as input, the attribute graph convolutional network learns attribute features using visual attention and graph convolutional network. Given $\mathbf{V}_r$ of $W \times H \times C$ dimension, it first utilizes two 1×1 convolution layers to produce a primitive attribute feature $\mathbf{X}_p$ with the size of $W \times H \times C_p$ ($7 \times 7 \times 128$). Three 1×1 convolution layers are then utilized to learn an initial attention map with size of $W \times H \times N$, where N denotes the number of attributes. Each channel $\boldsymbol{a}_n$ in the attention map corresponds to the activation response of a certain attribute (*e.g. blue shirt, long hair, brown shoes, black bag*). The attention map is then used to squeeze the primitive attribute feature into N attentive attribute feature vectors through a channel-wise convolution. Each attentive attribute feature vector is computed by:

$$\begin{aligned} \mathbf{x}_n^a(k) &= \sum_{i=0}^{W}\sum_{j=0}^{H} \boldsymbol{a}_n(i,j)\mathbf{X}_p(i,j,k), \\ k &= 1, \cdots, C_p; \quad n = 1, \cdots, N \end{aligned} \tag{1}$$

where $\boldsymbol{x}_n^a(k)$ denotes the k-th element of the attentive feature of n-th attribute.

The graph convolution network aims to exploit the underlying semantic context among attributes, towards improving attribute learning. Each attribute feature $\boldsymbol{x}_n^a$ corresponds to a node in the graph, and a scalar w_{nm} denotes the association weight between the n-th and m-th node. w_{nm} represents the magnitude of context between attributes and is computed as:

$$w_{nm} = \begin{cases} \frac{exp(S(\boldsymbol{x}_n^a, \boldsymbol{x}_m^a))}{\sum_{j \in \mathcal{N}(n)} exp(S(\boldsymbol{x}_n^a, \boldsymbol{x}_j^a))}, & n \neq m, \ m \in \mathcal{N}(n) \\ 0, \ else \end{cases} \tag{2}$$

in which $\mathcal{N}(n)$ is the set of 20 nearest neighbor nodes to the n-th node under Euclidean distance. The node $\boldsymbol{x}_n^a$ is updated by neighbor nodes as $[\boldsymbol{x}_n^a, \boldsymbol{x}_n^e]$, where $\boldsymbol{x}_n^e = \sum_{m \in \mathcal{N}(n)} w_{nm}\boldsymbol{x}_m^a$. The updated $\boldsymbol{x}_n^a$ then passes through a fully-connected layer to produce the feature of each attribute, *i.e.*, $\boldsymbol{A}_n$ of 64 dimension. Based on $\boldsymbol{A}_n$, a binary classifier with a cross-entropy loss function is used to predict the presence of each attribute.

By concatenating $\boldsymbol{A}_n$ of various attributes, we obtain the overall attribute feature $\boldsymbol{v}_a$. It is in turn fused with the appearance feature $\boldsymbol{v}_p$ to obtain the comprehensive visual representation as follows:

$$\boldsymbol{v} = ReLU(\boldsymbol{\omega}_p\boldsymbol{v}_p + \boldsymbol{\omega}_a\boldsymbol{v}_a) - (\boldsymbol{\omega}_p\boldsymbol{v}_p - \boldsymbol{\omega}_a\boldsymbol{v}_a)^2 \tag{3}$$

where $\boldsymbol{\omega}_p$ and $\boldsymbol{v}_a$ are the weighting parameters. This fusion strategy leads to faster convergence of training compared to the popular bi-linear fusion [60]. $\boldsymbol{v}_p$ and $\boldsymbol{v}_a$ are transferred to the same dimension of $1,024$ before fusion.

B. Hierarchical Text Embedding Network

As shown in Figure 4, the hierarchical text embedding network consists of a word embedding layer, multi-stacked bidirectional LSTMs (Bi-LSTMs) [61]–[63] and a textual attention block.

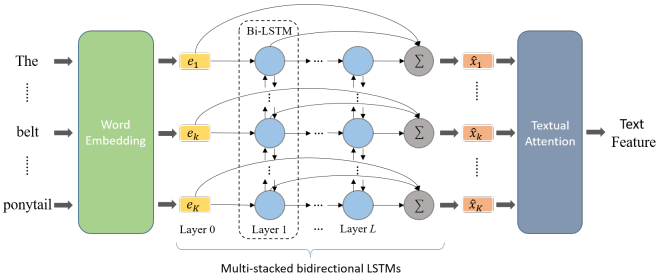

Fig. 4. Detailed structure of the hierarchical text embedding network.

Given a text description, we first use basic tokenizing and split it into words. Each word is represented as a one-hot vector, which is in turn sent to the word embedding layer to get a compact token representation $\boldsymbol{e}_m$, where $m = 1, \cdots, M$ and M is the number of words in the text description. The token representations are then passed through the multi-stacked bidirectional LSTMs to get contextual representation. The multi-stacked Bi-LSTMs includes L layers of Bi-LSTMs, each of 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 propagates relevant historical and future context to next word.

At each word position m, the output of the forward LSTM in each layer can be written as $\overrightarrow{\boldsymbol{h}}_{m,j}$ where $j = 1, \cdots, L$. The output of the backward LSTM in each layer is $\overleftarrow{\boldsymbol{h}}_{m,j}$. For each $\boldsymbol{e}_m$, the multi-stacked Bi-LSTMs outputs a set of $2L+1$ representations.

$$\begin{aligned} R_m &= \{\boldsymbol{e}_m, \overrightarrow{\boldsymbol{h}}_{m,j}, \overleftarrow{\boldsymbol{h}}_{m,j} | \ j = 1, \cdots, L\} \\ &= \{\boldsymbol{x}_{m,j} \mid j = 0, \cdots, L\} \end{aligned} \tag{4}$$

where $\boldsymbol{x}_{m,j}$ is the concatenation of $\overrightarrow{\boldsymbol{h}}_{m,j}$ and $\overleftarrow{\boldsymbol{h}}_{m,j}$, $\boldsymbol{x}_{m,0}$ is $\boldsymbol{e}_m$. In order to utilize the context within text description, we learn a linear aggregation of the representations $\boldsymbol{x}_{m,j}$ from all Bi-LSTM layers. The final representation of m-th word is:

$$\boldsymbol{x}_m = \sum_{j=0}^{L} s_j \boldsymbol{x}_{m,j} \tag{5}$$

in which s_j denotes the weights of softmax normalization.

Afterwards, we use a textual attention block to produce more precise text feature. The scaled dot-product attention mechanism [64] is utilized here. Based on the d_k-dimensional feature $\boldsymbol{x}_m$ of each word from multi-stacked Bi-LSTMs, the feature of the text query is a matrix $\mathbf{X} \in \mathbb{R}^{K \times d_k}$. The final

textual feature $\boldsymbol{t}$ is computed by $Attention(\mathbf{X})$ followed by a sum-pooling operation.

$$Attention(\mathbf{X}) = softmax(\frac{\mathbf{X}\mathbf{X}^T}{\sqrt{d_k}})\mathbf{X} \tag{6}$$

Note that the scaling factor $\frac{1}{\sqrt{d_k}}$ is used to prevent the inner product of two word representations to be too large.

C. Cross-modal Adversarial Learning

As aforementioned, there is a demand for discriminative multi-modal features with modal-invariant distribution. Inspired by the success of Generative Adversarial Network (GAN) [65] and adversarial learning [27]–[29], we develop a cross-modal adversarial learning module to learn discriminative and modality-invariant visual-textual representation. It consists of a modality discriminator and a cross-modal learner, as illustrated in Figure 5, which perform a min-max game in an adversarial learning manner. The modality discriminator is designed to distinguish the features from visual and textual modalities. It predicts the probability of whether the vsiual and textual features belonging to two different modalities. An adversarial loss function is used to classify the modality label (*i.e.*, the same or different) as follows:

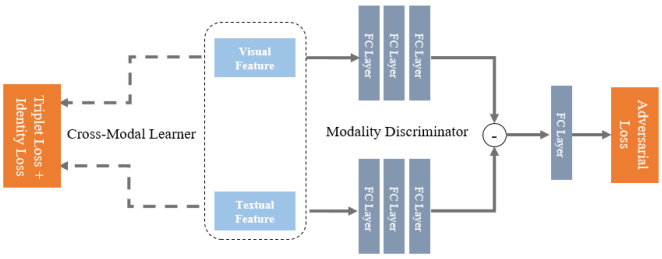

Fig. 5. Detailed structure of the cross-modal adversarial learning module, consisting of a modality discriminator and a cross-modal learner.

$$\mathcal{L}_{adv}(\boldsymbol{\theta}_D) = -\frac{1}{N}\sum_{i=1}^{N}(d_i \log P_i(\boldsymbol{\theta}_D) + (1-d_i)(1-\log P_i(\boldsymbol{\theta}_D))) \tag{7}$$

in which $\boldsymbol{\theta}_D$ denotes the model parameters. $P_i(\boldsymbol{\theta}_D)$ refers to the classification probability of the i-th pair of image and text. $d_i = 1$ for a positive pair and $d_i = 0$ otherwise.

A cross-modal learner is used to improve the discriminativeness of visual-textual representation as well as preserve the semantic affinity between the image and text of the same pedestrian. It uses an identification loss and a triplet loss. The former encourages the visual and textual features to be discriminative, the latter preserves the semantic affinity. The identification loss is formulated as follows:

$$\mathcal{L}_{ide}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathbf{W}_{y_i}^T \boldsymbol{z}_i + \boldsymbol{b}}}{\sum_{j=1}^{K} e^{\mathbf{W}_j^T \boldsymbol{z}_j + \boldsymbol{b}}} \tag{8}$$

where y_i is the person ID of the i-th sample (*i.e.*, image or text), and $\boldsymbol{z}_i$ refers to the visual feature $\boldsymbol{v}$ or textual feature $\boldsymbol{t}$ of the sample. $\boldsymbol{\theta}_V$ and $\boldsymbol{\theta}_T$ denote the parameters of the visual attribute graph network and hierarchical text embedding network, respectively. $\mathbf{W}_j \in \mathbb{R}^{1,024}$ is the j-th column of the weight matrix $\mathbf{W} \in \mathbb{R}^{1,024\times 11,003}$ and $\boldsymbol{b}$ is a bias term. The size of $\mathbf{W}$ is related to the feature dimension and the number of pedestrians in query gallery.

The triplet loss optimizes the feature embedding space in which samples with the same identity are closer to each other than those with different identities, as follows:

$$\begin{aligned}\mathcal{L}_{tri}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) = &\frac{1}{N}\sum_{i=1}^{N} max(0, m + d(\boldsymbol{v}_i, \boldsymbol{t}_i^+) - d(\boldsymbol{v}_i.\boldsymbol{t}_i^-)) \\ &+\frac{1}{N}\sum_{j=1}^{N} max(0, m + d(\boldsymbol{t}_j, \boldsymbol{v}_j^+) - d(\boldsymbol{t}_j, \boldsymbol{v}_j^-,)),\end{aligned} \tag{9}$$

where $\boldsymbol{v}, \boldsymbol{t}$ are the visual and textual features, respectively. $\boldsymbol{t}_i^+, \boldsymbol{t}_i^-$ correspond to the positive and negative text descriptions of i-th image. $\boldsymbol{v}_j^+, \boldsymbol{v}_j^-$ correspond to the positive and negative images of j-th text description. $d(\cdot,\cdot)$ is the Euclidean distance. m is the margin and set empirically in the experiments. The overall loss function of the cross-modal learner can be written as:

$$\mathcal{L}_{xs}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) = \mathcal{L}_{tri} + \alpha \cdot \mathcal{L}_{ide} \tag{10}$$

The model training is based on adversarial min-max optimization as similar to GAN:

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_V, \hat{\boldsymbol{\theta}}_T) &= \arg\min_{\boldsymbol{\theta}_V, \boldsymbol{\theta}_T} \ (\mathcal{L}_{xs}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) - \mathcal{L}_{adv}(\boldsymbol{\theta}_D)) \\ \hat{\boldsymbol{\theta}}_D &= \arg\max_{\boldsymbol{\theta}_D} \ (\mathcal{L}_{xs}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) - \mathcal{L}_{adv}(\boldsymbol{\theta}_D))\end{aligned} \tag{11}$$

IV. Experiments

In this section, we conduct extensive experiments to evaluate the proposed AATE on a challenging text-based person search benchmark *i.e.*, CUHK-PEDES and compare it to state-of-the-art methods. We also conduct experiments on a cross-modal retrieval dataset *i.e.*, Caltech-UCSD birds (CUB-200-2011). The experimental results show that AATE achieves superior performance over the state-of-the-art methods. Moreover, we investigate the effectiveness of each component of AATE, including the visual attribute graph convolutional network, the hierarchical text embedding network and the cross-modal adversarial learning module.

Datasets- CUHK-PEDES [15] is the only dataset at the moment for person search with natural language query. It contains 40,206 pedestrian images of 13,003 identities. Each pedestrian image is described by two textual descriptions. The dataset is splitted into three subsets for training, validation, and test, respectively, with non-overlapping identities. The training subset contains 11,003 identities with 34,054 images and 68,108 text descriptions. The validation subset contains 1,000 identities with 3,078 images, while the test subset has 3,074 images of the remaining 1,000 identities. The CUB-200-2011 dataset contains 11,788 images of 200 bird categories. Each image is labeled with ten text descriptions. The dataset

TABLE I
PERFORMANCE COMPARISON TO THE STATE-OF-THE-ART METHODS ON THE CUHK-PEDES DATASET.

Method	Text-to-Image		
	Top-1 (%)	**Top-5** (%)	**Top-10**(%)
CNN-RNN [68]	8.07	-	32.47
Neural Talk [69]	13.66	-	41.72
GNA-RNN [15]	19.05	-	53.64
IATVM [19]	25.94	-	60.48
PWM-ATH [16]	27.14	49.45	61.02
GLA [20]	43.58	66.93	76.26
CAN [21]	45.52	67.12	76.98
CMPM-CMPC [14]	49.37	-	79.27
A-GANet [22]	**53.14**	74.03	81.95
AATE	52.42	**74.98**	**82.74**

TABLE II
PERFORMANCE COMPARISON TO THE STATE-OF-THE-ART METHODS ON THE CUB-200-2011 DATASET.

Method	Image-to-Text	Text-to-Image
	Top-1(%)	**AP@50**(%)
Bow [70]	44.1	39.6
Word2Vec [71]	38.6	33.5
Word CNN [68]	51.0	43.3
Word CNN-RNN [68]	56.8	48.7
GMM+HGLMM [72]	36.5	35.6
Triplet [19]	52.5	52.4
Latent Co-attention [19]	61.5	57.6
CMPM+CMPC [14]	64.3	67.9
AATE	**65.8**	**71.5**

is splited as 100 categories for training, 50 for validation, and 50 for testing.

Implementation Details- The implementation of the proposed method is based on the Tensorflow framework with four NVIDIA Titan XP GPUs. The Adam optimization is started with learning rate lr of 0.0002, the weight decay of $5e^{-4}$. The parameter α in the cross-modal learning loss is set to 1. All the images are resized to the size of $224 \times 224 \times 3$ and normalized with $1/256$. The training set is augmented by random horizontal flipping. The number of mini-batches is set to 16. The proposed model is optimized for 50 epochs in total. For CUHK-PEDES dataset, we utilize the Stanford CoreNLP toolkit [66] to select noun phrases in the text descriptions and obtain a set of attributes. We then filter out the attributes which appear in less than 300 pedestrians and manually select 100 attributes for experiments. The CUB-200-2011 dataset provides 312 attributes of birds. The word embedding matrix in the hierarchical text embedding network on the CUHK-PEDES and CUB-200-2011 datasets are with the size of $3,600 \times 512$ and $2,500 \times 512$, respectively. The size of Bi-LSTM is set to 512, and the number of Bi-LSTM layers L is set to 2. The visual and textual features are of 1,024 dimension. We use the pre-trained ResNet-50 model on ImageNet dataset [67] as the initialization of the base network. We first train the attribute graph convolutional network for attribute recognition and the optimize the entire AATE for text-based person search.

Protocol- For person search task on the CUHK-PEDES dataset, we report the performance of searching pedestrian images with text query, *i.e.*, the Top-K accuracy. It indicates the fraction of queries for which at least one relevant image is retrieved in the closest K samples to the query. For the cross-modal retrieval task on CUB-200-2011 dataset, we adopt Average Precision at 50 ($AP@50$) as the performance metric for text-to-image retrieval and Top-1 accuracy for image-to-text retrieval, following the previous works [14], [19]. $AP@50$ indicates the percent of top-50 retrieved images whose class matches that of the text query, averaged over all the test classes.

A. Comparison to State-of-the-Arts

CUHK-PEDES: We compare the proposed AATE to nine state-of-the-art methods on the CUHK-PEDES dataset, including the similarity learning networks CNN-RNN [68], NeuralTalk [69], GNA-RNN [15], IATVM [19], PWM-ATH [16] and GLA [20], as well as the feature embedding methods CAN [21], CMPM-CMPC [14] and A-GANet [22]. TABLE I shows the top-1, top-5, and top-10 accuracy of person search via text query. The proposed AATE achieves 52.42%, 74.98%, 82.74% of top-1, top-5 and top-10 accuracy, respectively. We can see that AATE achieves the best top-5 and top-10 accuracy and outperforms almost all the compared methods in terms of top-1 accuracy. Moreover, AATE improves the performance by a large margin compared to existing approaches of similarity learning networks. The results demonstrate the effectiveness of AATE on the task of person search by text description.

CUB-200-2011: TABLE II reports the performance of image-to-text and text-to-image retrieval on the CUB-200-2011 dataset. We compare AATE to eight state-of-the-art approaches, including BoW [70], Word2Vec [71], Word CNN [68], Word CNN-RNN [68], GMM+HGLMM [72], Triplet [19], Latent Co-attention [19], and CMPM-CMPC [14]. We can see that AATE obtains the best performance, in particular, 65.8% top-1 accuracy for image-to-text retrieval and 71.5% AP@50 for text-to-image retrieval. AATE outperforms the second best method, *i.e.*, CMPM-CMPC, by 1.5% in terms of top-1 accuracy and 3.6% in terms AP@50, respectively. The experimental results demonstrate the capacity of AATE in visual-textual representation learning.

B. Ablation Studies

To investigate the effectiveness and contribution of the components of AATE, we conduct a series of ablation experiments on the CUHK-PEDES dataset.

TABLE III reports the leave-one-out evaluation results of AATE. AATE_\attr is the variant of AATE which is without the visual attribute graph network and only uses the appearance feature. AATE_\hier refers to the variant that uses a single-layer Bi-LSTM instead of the multi-stacked Bi-LSTMs

Fig. 6. Visualization of the spatial attention corresponding to pedestrian attributes.

and textual attention block. AATE_\adv refers to the variant without the cross-modal adversarial learning module. It only uses the identification loss and triplet loss. From the results, we can see that AATE_\attr, AATE_\hier and AATE_\adv give rise to performance degradation compared to AATE in terms of all the top-1, top-5 and top-10 accuracy. For example, the top-1 accuracy of them have 1.94%, 1.1% and 1.44% drops compared to the performance of AATE, respectively. This indicates the effect of the three components in boosting the performance.

Analysis of the attribute graph convolutional network. We investigate the effect of the visual attribute graph in AATE. AATE_\gcn is the variant of AATE that has the attribute attention block but not the graph convolution module. TABLE IV reports the performance comparison among AATE_\attr, AATE_\gcn and AATE. We can observe that AATE_\gcn has performance improvement over AATE_\attr but performance degradation compared to AATE, in terms of all the top-1, top-5 and top-10 accuracy. The improvement over AATE_\attr demonstrate the usefulness of visual attributes and the effectiveness of attribute feature. The performance improvements of AATE over AATE_\gcn indicates that the attribute graph convolution module is able to learn more effective attribute features by exploiting semantic context among attributes.

Analysis of the hierarchical text embedding network. We conduct experiments to investigate the effect of the multi-stacked Bi-LSTMs and textual attention block within the hierarchical text embedding network. The experimental results are provided in TABLE V. AATE_\msbl refers to a variant of AATE using a single layer of Bi-LSTM instead of the multi-stacked Bi-LSTMs. AATE_\attn refers to a variant without the text attention block. TABLE V provides the experimental results, from which we can see that either AATE_\msbl or AATE_\attn leads to performance degradation compared to AATE. This indicate the effect of the multi-stacked Bi-LSTMs and textual attention in learning textual features.

Analysis of the cross-modal adversarial learning module. The cross-modal adversarial learning is based on three types of loss functions, including the adversarial loss $\mathcal{L}_{adv}$, identification loss $\mathcal{L}_{ide}$ and triplet loss $\mathcal{L}_{tri}$. We investigate

TABLE III
EVALUATION OF THE EFFECTIVENESS OF THE COMPONENTS WITHIN AATE ON THE CUHK-PEDES DATASET.

Method	Text-to-image		
	Top-1(%)	**Top-5**(%)	**Top-10**(%)
AATE_\attr	50.48	73.82	82.44
AATE_\hier	51.32	74.15	82.61
AATE_\adv	50.98	73.84	82.52
AATE	52.42	74.98	82.74

TABLE IV
EVALUATION OF THE EFFECTIVENESS OF THE COMPONENTS WITHIN THE ATTRIBUTE GRAPH CONVOLUTION NETWORK ON THE CUHK-PEDES DATASET.

Method	Text-to-image		
	Top-1(%)	**Top-5**(%)	**Top-10**(%)
AATE_\attr	50.48	73.82	82.44
AATE_\gcn	51.40	74.45	82.46
AATE	52.42	74.98	82.74

TABLE V
EVALUATION OF THE EFFECTIVENESS OF THE COMPONENTS WITHIN THE HIERARCHICAL TEXT EMBEDDING NETWORK ON CUHK-PEDES DATASET.

Method	Text-to-image		
	Top-1(%)	**Top-5**(%)	**Top-10**(%)
AATE_\msbl	51.57	74.52	82.65
AATE_\attn	51.94	74.66	82.70
AATE	52.42	74.98	82.74

TABLE VI
EVALUATION OF THE EFFECTIVENESS OF THE LOSS FUNCTIONS WITHIN THE CROSS-MODAL ADVERSARIAL LEARNING MODULE ON CUHK-PEDES DATASET.

Method	Text-to-image		
	Top-1(%)	**Top-5**(%)	**Top-10**(%)
AATE_$\mathcal{L}_{ide}$	34.29	59.23	69.35
AATE_$\mathcal{L}_{tri}$	47.67	70.83	80.13
AATE_$\mathcal{L}_{xs}$	50.98	73.84	82.52
AATE	52.42	74.98	82.74

the effect of each loss function by evaluating several variants of AATE. AATE_$\mathcal{L}_{ide}$ refers to the variant only using the identification loss. AATE_$\mathcal{L}_{tri}$ is the one only using the triplet loss. AATE_$\mathcal{L}_{xs}$ uses the identification and triplet loss. TABLE VI reports the experimental results. We can see that AATE_$\mathcal{L}_{xs}$ obtains better accuracy than AATE_$\mathcal{L}_{tri}$ and AATE_$\mathcal{L}_{ide}$. This indicates that the combination use of identification and triplet loss can better deal with the inter-class invariance and intra-class variance, and leads to more effective features. AATE achieves the best performance. The improvement of AATE over AATE_$\mathcal{L}_{xs}$ demonstrate the effectiveness of the cross-modal adversarial learning. It is able to improve the discriminativeness of visual-textual features as well as the modal-invariance of feature distributions across different modalities.

Qualitative evaluation. We conduct qualitative evaluation for the proposed AATE. Figure 7 (a) shows the qualitative results of person search with text description by AATE on CUHK-PEDES dataset. For each text query, we show the top-10 retrieved images. We can see that each retrieved image has multiple regions that match the description. Although some irrelevant images are retrieved, they has regions relevant to part of the description. Figure 7 (b) shows the qualitative results of cross-modal retrieval on CUB-200-2011 dataset. Moreover, we inspect whether AATE can concentrate on corresponding regions when learning attributes. Figure 6 visualized the spatial attention inferred by AATE for some attributes. We can see that AATE can attend to the corresponding regions, leading to precise attribute features.

V. CONCLUSIONS

In this work, we proposed a novel adversarial attribute-text embedding network (ATTE) for person search with natural language query. ATTE consists of an attribute graph neural network, a hierarchical text embedding network and a cross-modal adversarial learning module. The cross-modal adversarial learning module was proposed to learn discriminative and modality-invariant image-text representation for cross-modal matching, through a min-max game between a cross-modal learner and a modality discriminator. The attribute graph neural network learns visual attributes precisely by jointly exploiting visual attentions on attributes and the semantic dependencies among attributes. The hierarchical text embedding network exploits context in natural language description and concentrates on informative words, learning effective textual feature. We conducted extensive experiments to evaluate ATTE on two challenging benchmarks, *i.e.*, CUHK-PEDES and Caltech-UCSD birds datasets. The experimental results have demonstrated the effectiveness of the proposed approach.

ACKNOWLEDGMENT

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, U19B2038 and 61620106009 as well as the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

REFERENCES

[1] 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 *Proceedings of the European Conference on Computer Vision*, 2018, pp. 480–496.

[2] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2119–2128.

[3] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 369–378.

[4] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "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*, 2018, pp. 420–429.

[5] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7202–7211.

[6] J. Liu, Z.-J. Zha, H. Xie, Z. Xiong, and Y. Zhang, "Ca3net: Contextual-attentional attribute-appearance network for person re-identification," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 737–745.

[7] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[8] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[9] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1470–1478.

[10] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1983–1991.

Fig. 7. (a) Examples of top-10 retrieved images with text query by the proposed AATE on the CUHK-PEDES dataset. (b) Examples of top-10 retrieved images with text query by AATE on the CUB-200-2011 dataset. Relevant images are marked with red rectangles.

[11] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "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*, 2018, pp. 1169–1178.

[12] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4733–4742.

[13] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2429–2438.

[14] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 686–701.

[15] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1970–1979.

[16] T. Chen, C. Xu, and J. Luo, "Improving text-based person search by spatial matching and adaptive threshold," in *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1879–1887.

[17] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 201–216.

[18] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen, "Dual-path convolutional image-text embedding with instance loss," *arXiv preprint arXiv:1711.05535*, 2017.

[19] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1890–1899.

[20] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang, "Improving deep visual representation for person re-identification by global and local image-language association," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 54–70.

[21] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Cascade attention network for person search: Both image and text-image similarity selection," *arXiv preprint arXiv:1809.08440*, 2018.

[22] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, "Deep adversarial graph attention convolution network for text-based person search," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, pp. 665–673.

[23] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5005–5013.

[24] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4107–4116.

[25] M. Dorfer, J. Schlüter, A. Vall, F. Korzeniowski, and G. Widmer, "End-to-end cross-modality retrieval with cca projections and pairwise ranking loss," *International Journal of Multimedia Information Retrieval*, vol. 7, no. 2, pp. 117–128, 2018.

[26] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3441–3450.

[27] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7181–7189.

[28] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen, "Unsupervised cross-modal retrieval through adversarial learning," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 1153–1158.

[29] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 154–162.

[30] M. Yamaguchi, K. Saito, Y. Ushiku, and T. Harada, "Spatio-temporal person retrieval via natural language queries," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1453–1462.

[31] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3668–3678.

[32] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1219–1228.

[33] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7239–7248.

[34] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2360–2367.

[35] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the 10th European Conference on Computer Vision*. Springer, 2008, pp. 262–275.

[36] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proceedings of the 12th international conference on Computer Vision*. Springer, 2012, pp. 413–422.

[37] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[38] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," *arXiv preprint arXiv:1604.01850*, vol. 1, no. 2, 2016.

[39] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *Proceedings of the British Machine Vision Conference*, vol. 2, no. 5, 2010, p. 6.

[40] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 649–656.

[41] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2288–2295.

[42] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3318–3325.

[43] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2990–2999.

[44] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3908–3916.

[45] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[46] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei, "Multi-scale triplet cnn for person re-identification," in *Proceedings of the ACM Conference on Multimedia Conference*. ACM, 2016, pp. 192–196.

[47] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1325–1334.

[48] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4747–4756.

[49] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1345–1353.

[50] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2501–2514, 2016.

[51] X. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5683–5695, 2018.

[52] J. Liu, Z.-J. Zha, X. Chen, Z. Wang, and Y. Zhang, "Dense 3d-convolutional neural network for person re-identification in videos," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1s, pp. 1–19, 2019.

[53] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proceeding of the International Joint Conference on Artificial Intelligence*, 2017, pp. 2194–2200.

[54] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 486–504.

[55] W. Zhang, S. Hu, K. Liu, and Z.-J. Zha, "Learning compact appearance representation for video-based person re-identification," *IEEE Transac-*

tions on Circuits and Systems for Video Technology, vol. 29, no. 8, pp. 2442–2452, 2019.
[56] Z. Meng, N. Adluru, H. J. Kim, G. Fung, and V. Singh, "Efficient relative attribute learning using graph neural networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 552–567.
[57] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, and T.-S. Chua, "Robust (semi) nonnegative graph embedding," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 2996–3012, 2014.
[58] H. Zhang, Z.-J. Zha, S. Yan, J. Bian, and T.-S. Chua, "Attribute feedback," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 79–88.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[60] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," *arXiv preprint arXiv:1802.05766*, 2018.
[61] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 299–307.
[62] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.
[63] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *Proceedings of the ACM International Conference on Multimedia*, 2016.
[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
[65] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of Advances in Neural Information Processing System*, 2014, pp. 2672–2680.
[66] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60.
[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
[68] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58.
[69] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
[70] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.
[71] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.
[72] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4437–4446.

Zheng-Jun Zha (M'08) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2004 and 2009, respectively. He was a Researcher with the Hefei Institutes of Physical Science, Chinese Academy of Sciences, from 2013 to 2015, a Senior Research Fellow with the School of Computing, National University of Singapore (NUS), from 2011 to 2013, and a Research Fellow with NUS from 2009 to 2010. He is currently a Full Professor with the School of Information Science and Technology, University of Science and Technology of China, the Executive Director of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application.

His research interests include multimedia analysis, retrieval and applications, and computer vision. He has authored or coauthored more than 100 papers in these areas with a series of publications on top journals and conferences. He was a recipient of multiple paper awards from prestigious multimedia conferences, including the Best Paper Award and Best Student Paper Award in ACM Multimedia and so on. He serves as an Associate Editor for the IEEE Transactions on Circuits and System for Video Technology.

Jiawei Liu received the B.E. degree from Hefei University of Technology, China, in 2013 and received the Ph.D. degree from the University of Science and Technology of China, China, in 2019. He is currently a postdoctoral fellow at the School of Information Science and Technology, University of Science and Technology of China. His research interests mainly include computer vision and multimedia.

Di Chen received the B.E. degree from Dalian University of Technology, China, in 2016 and received the M.S. degree from the University of Science and Technology of China, China, in 2019. His research interests mainly include computer vision and multimedia.

Feng Wu Feng Wu (M'99-SM'06-F'13) received the B.S. degree in Electrical Engineering from Xidian University in 1992, and received the M.S. and Ph.D. degrees in Computer Science from the Harbin Institute of Technology in 1996 and 1999, respectively. He was a Principle Researcher and Research Manager with Microsoft Research Asia, Beijing, China. He is currently a Professor and the Dean of the School of Information Science and Technology, University of Science and Technology of China. He has authored or co-authored more than 300 high-quality papers, including several dozens of IEEE Transactions papers and top conference papers on MOBICOM, SIGIR, CVPR, and ACM MM. He has 77 granted US patents. His research interests include image and video compression, media communication, and media analysis and synthesis. Dr. Wu serves or had served as an Associate Editor of IEEE Transactions on Circuits and System for Video Technology, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, and several other international journals. He also serves as the TPC Chair of MMSP 2011, VCIP 2010, and PCM 2009, and the Special Sessions Chair of ICME 2010 and ISCAS 2013. He received the best paper awards in IEEE TCSVT 2009, PCM 2008 and VCIP 2007, as well as the IEEE Circuits and Systems Society 2012 Best Associate Editor Award.