Person Reidentification via Structural Deep Metric Learning

Xun Yang, Peicheng Zhou, and Meng Wang, *Senior Member, IEEE*

***Abstract*— Despite the promising progress made in recent years, person reidentification (re-ID) remains a challenging task due to the complex variations in human appearances from different camera views. This paper proposes to tackle this task by jointly learning feature representation and distance metric in an end-to-end manner. Existing deep metric learning-based re-ID methods usually encounter the following two weaknesses: 1) most works based on pairwise or triplet constraints often suffer from slow convergence and poor local optima, partially because they use very limited samples for each update and 2) hard negative sample mining has been widely applied in existing works. However, hard positive samples, which also contribute to the training of network, have not received enough attention. To alleviate these problems, we develop a novel structural metric learning objective for person re-ID, in which each positive pair is allowed to be compared against all negative pairs in a minibatch and each positive pair is adaptively assigned a hardness-aware weight to modulate its contribution. The introduced positive pair weighting strategy enables the algorithm to focus more on the hard positive samples. Furthermore, we propose to enhance the proposed loss function by adding a global loss term to reduce the variances of positive/negative pair distances, which is able to improve the generalization capability of the network model. By this approach, person images can be nonlinearly mapped into a low-dimensional embedding space where similar samples are kept closer and dissimilar samples are pushed farther apart. We implement the proposed algorithm using the inception architecture and evaluate it on three large-scale re-ID data sets. Experiment results demonstrate that our approach is able to outperform most state of the arts while using much lower dimensional deep features.**

***Index Terms*— Computer vision, deep metric learning, deep neural network, machine learning, person re-identification (re-ID).**

I. Introduction

IN RECENT years, person reidentification (re-ID) [1]–[4] has attracted increasing attention in the computer vision community for its critical role in security surveillance applications. It aims to recognize the person of interest across multiple nonoverlapping camera views. Given a probe person image (query), the task is to rank all the person images in the gallery set by the similarity between the query and candidate images and return the most relevant images as retrieval results. To tackle this problem, massive efforts [5]–[20] have been made over the last decade. However, it remains a challenging problem since a person's appearance usually undergoes dramatic variations across camera views due to the changes in view angle, body pose, illumination, and background clutter.

Traditional methods mainly consist of two parts: feature extraction and metric learning. The first part focuses on designing robust hand-crafted features [6]–[10]. The second part [5], [11]–[17], [21], [22] aims to learn a suitable distance/similarity function. Despite the promising progress, they optimize the two parts either separately or sequentially, which may result in a suboptimal performance. Once useful information has been lost in the feature extraction stage, it can hardly be recovered later.

More recently, deep convolutional neural networks (CNNs) have gained increasing popularity in person re-ID [3], [4], [23]–[41]. Different from traditional works, deep CNN-based approaches are able to learn feature representation and distance metric jointly in an end-to-end manner, where the major task is to learn a nonlinear discriminative mapping from person images to low-dimensional embeddings, where similar examples are mapped close to each other, while dissimilar examples are pushed farther apart.

Existing deep CNN-based re-ID approaches can roughly be classified into two groups. The first group [26]–[30], [42] considers each sample independently using an identification loss, which directly casts person re-ID as a multiclass recognition task and usually learns a nonlinear mapping from an input person image to its person identity using a cross-entropy loss. Despite the simplicity, the first group of methods may be less effective on small data sets, and the identification loss usually suffers from the large intrapersonal variance. The second group [25], [31]–[40] prefers to learn discriminative feature embeddings by employing pairwise or triplet constraints. Pairwise constraint-based methods [25], [31]–[36] take paired person images (positive/negative pairs) as an input to minimize a verification/contrastive loss. They usually focus on both reducing interpersonal variations and enlarging interpersonal variations and result in an absolute distance. Triplet constraint-based methods [37]–[40] take image triplets for each update and minimize a triplet loss that encourages the network to find an embedding space where the distance

Manuscript received July 19, 2017; revised March 26, 2018 and July 22, 2018; accepted July 25, 2018. Date of publication August 24, 2018; date of current version September 18, 2019. This work was supported by the National Nature Science Foundation of China under Grant 61732008 and Grant 61725203. *(Corresponding author: Meng Wang.)*

X. Yang and M. Wang are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: hfutyangxungmail.com; eric.mengwanggmail.com).

P. Zhou is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2861991

between positive pair (*anchor*, *positive*) is smaller than that between negative pair (*anchor*, *negative*) by a fixed margin. Triplet constraints result in a relative distance which usually matters more than absolute distance for most tasks [43]–[47], e.g., face verification, image clustering, and retrieval. The second group has received much more attention in recent years. This paper is more related to the second group.

Despite yielding promising progress, pairwise or triplet constraint-based works often suffer from slow convergence and poor local optima [46], partially because they only employ two samples (pairwise constraints) or three samples (triplet constraints) for each update. Moreover, most existing works focus more on mining hard negative samples for the optimization of network, while they pay less attention to hard positive samples that also contribute to the optimization.

To alleviate these problems, this paper proposes to learn feature representation and distance metric jointly in an end-to-end manner for person re-ID with a hardness-aware structural metric learning objective. The proposed learning objective improves the triplet loss by allowing each positive pair to be compared against all the corresponding negative pairs within minibatch for each update. To better leverage hard positive samples, each positive pair is adaptively assigned a hardness-aware weight in the proposed learning objective, which enables the algorithm to focus more on hard positive samples rather than treat all positive pairs equally. By this way, the learning performance can be effectively improved. In addition, the second-order statistics of positive/negative pair distances is incorporated into our learning objective to improve the generalization capability of the network. Specifically, we introduce a global loss term that penalizes the large variances of positive/negative pair distances. By the proposed approach, we expect to learn more discriminative deep features that are robust to person appearance variation. Extensive experiments on several large-scale data sets have demonstrated the effectiveness of the proposed structural feature embedding learning approach.

II. Related Work

A. Deep Metric Learning

Distance metric learning [48], [49] aims to learn a mapping from the input data space to a low-dimensional embedding space, where similar instances are kept closer, while dissimilar instances are pushed farther apart. In recent years, with the remarkable development of deep learning techniques [50]–[54], deep metric learning has shown promising results on multiple computer vision tasks, e.g., face recognition, image retrieval, and fine-grained image recognition. The major difference with standard metric learning is that deep metric learning optimizes feature and metric jointly in an end-to-end manner. Existing works can be roughly classified into the following three groups.

The first group of deep metric learning methods trains Siamese networks with a contrastive loss [55]–[57], in which paired data are fed into neural networks. They usually minimize intraclass distance and penalize interclass distance for being smaller than a data-independent threshold. The contrastive loss usually results in the absolute distance.

The second group of methods aims to learn deep embeddings using the triplet loss [43], [44], [58], which takes triplets as an input. Each triplet consists of three samples (*anchor*, *positive*, and *negative*), where the former two samples share the same class label and the third one is from a different class. Triplet loss encourages the network to find an embedding space where the anchor sample is closer to the positive sample than the negative sample. The triplet loss results in a relative distance that has been shown to be better than an absolute distance in most tasks. For the triplet-loss-based methods, it is crucial to mine hard samples, e.g., the semihard or hardest negative samples, to select triplets violating the triplet constraints for fast convergence.

The third group of methods focuses on improving the performance by exploiting more negative samples for each update [45], [46] or exploiting the global structure of embedding space [47], [59]. Song *et al.* [47] proposed a lifted structured embedding loss by lifting the vector of pairwise distances within the minibatch to the dense matrix of pairwise distances. Sohn [46] designed a *N-pair* loss that optimizes the log probability of identification loss directly, which is a special case of the lifted structured loss [45]. Then, a structured prediction framework [47] is applied to ensure that the score of the ground-truth clustering assignment is higher than the score of any other clustering assignment, which exploits the global structure of embedding space and results in an impressive performance.

Our proposed loss function can be seen an improvement of lifted structured loss [45] and *N-pair* loss [46]. However, our loss can directly process the ℓ_2-normalized features. It is an important architecture design in our model. Usually, an ℓ_2-normalization layer is not adopted in most state-of-the-art deep metric learning methods, since it bounds the Euclidean distance in a small range and makes the optimization difficult. Besides, the lifted structured loss and *N-pair* loss [46] treat all positive pairs equally. We argue that hard positive pairs contribute more to the training of network than easy positive pairs. For this reason, we improve the lifted structured loss by adaptively assigning larger weights to hard positive pairs. Our loss function also includes a global loss term that minimizes the variances of positive/negative pair distances. It can improve the generalization capability of the network. By this way, the learning performance can be further improved.

B. Person Reidentification

This paper focuses on tackling person re-ID with the proposed deep metric learning scheme. We roughly categorize most existing works into two types: traditional methods and *deep* model-based methods. In this section, we only briefly introduce some representative works.

Traditional methods usually cope with two subproblems, i.e., feature learning and metric learning, separately. Some works focus on designing hand-crafted features [6]–[8], [10] that are expected to be robust to complex variations in human appearances from different camera views, e.g., local maximal occurrence feature [7] and Gaussian of Gaussian [6]. Some works focus on learning an optimal distance/similarity

function [11]–[13], [15]–[17], [60]–[62] using hand-crafted features to better characterize the similarity between a pair of person images. Zheng *et al.* [62] formulated re-ID as a relative distance learning problem by maximizing the probability that relevant samples have a smaller distance than the irrelevant ones. Kostinger *et al.* [14] developed a simple and effective metric learning method by computing the difference between the intraclass and the interclass covariance matrix. As an improvement, Liao *et al.* [7] proposed a cross-view quadratic discriminant analysis method by learning a more discriminative distance metric and a low-dimensional subspace simultaneously. Generally, traditional re-ID models can be easily trained and have shown a promising performance on small data sets, e.g. VIPeR [63].

Recently, more researchers in re-ID [3], [4], [23], [26]–[41], [64], [65] prefer to learn feature and metric jointly in an end-to-end manner by training deep neural networks. For the *deep* re-ID methods, an identification loss is the first choice, which has been exploited in multiple pioneering works [26]–[29]. For example, Xiao *et al.* [29] proposed to jointly handle person detection and identification in an end-to-end framework. Verification or contrastive loss functions are also widely used [31]–[36], which are usually employed to train Siamese-like networks [31], [32] using paired person images as an input. Ahmed *et al.* [35] improved the Siamese model by computing the cross-input neighborhood difference, which can capture local relationships between two input images. Later, Varior *et al.* [33] incorporated long short-term memory modules into the Siamese network, which can process person image parts sequentially so that the spatial connections can be memorized. More recently, triplet-based losses have attracted more attention in re-ID [37]–[40] for its great success in face recognition (FaceNet [44]). Compared with pairwise constraint-based losses, triplet-based loss functions take triplets as an input and result in a relative distance. Some works [3], [41], [66] also proposed to combine multiple loss functions to further enhance the performance. For example, Wang *et al.* [3] presented a deep learning framework to jointly optimize single-image representation and cross-image representation for re-ID, which achieves a satisfying performance. Both pairwise and triplet constraints are exploited in [3].

We propose a *deep* person re-ID scheme with a hardness-aware structural metric learning objective. It allows each positive pair to be compared against all negative pairs within minibatch and can adaptively assign large weights to hard positive pairs.

III. Proposed Approach

This paper addresses the problem of person re-ID by training a deep convolutional network discriminatively. We aim to find a suitable distance function between two person images $\mathbf{x}_i$ and $\mathbf{x}_j$, which is expected to be small if $\mathbf{x}_i$ and $\mathbf{x}_j$ are from the same class or large if they are from different classes. In this paper, it is defined as a squared Euclidean distance between deep embeddings of person images: $d^2(\mathbf{x}_i, \mathbf{x}_j) = \|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2^2$, where $f_\theta(\cdot)$ is a nonlinear feature mapping parameterized by the network parameters θ (weight matrices, bias vectors, and so on). Therefore, the key step of this problem is to learn the nonlinear feature mapping function $f_\theta(\cdot)$ that can transform a person image $\mathbf{x}_i$ to a low-dimensional and discriminative embedding $f_\theta(\mathbf{x}_i) \in \mathbb{R}^d$ in the Euclidean space. We expect to directly output ℓ_2-normalized deep embeddings in the training stage rather than only normalize the embeddings in the testing stage. As shown in Fig. 1, an ℓ_2-normalization layer is added after the fully connected layer. During testing, we use the ℓ_2-normalized low-dimensional embeddings to compute the distance between a query person image (probe) and candidate person images (galleries) in the database. If a true match to the probe exists in the database, it should have a small distance with the probe and will be top-ranked. To learn $f_\theta(\cdot)$, the task is to design a discriminative loss function to supervise the training of neural network. For simplicity, in this section, we omit θ from $f_\theta(\cdot)$ and use d_{ij}^2 to replace $d^2(\mathbf{x}_i, \mathbf{x}_j)$.

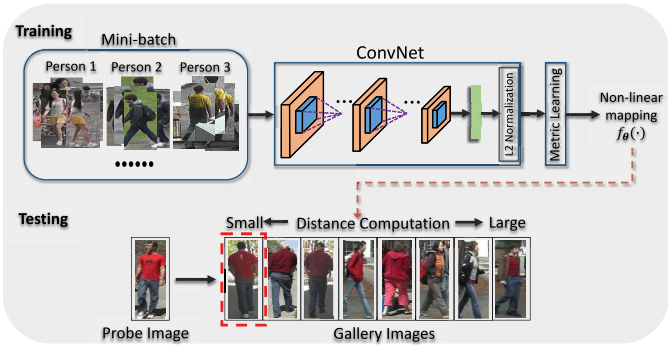

Fig. 1. Overview of the proposed person re-ID framework. The basic idea is to learn a nonlinear mapping from person images to discriminative embeddings based on a deep CNN. For each iteration, in the training stage, we feed the network with an identity-balanced minibatch generated by an online random sampling strategy and generate triplets online. The output of the last fully connected layer is ℓ_2-normalized and passed into the loss layer. The network parameters are updated by backpropagation supervised by a hardness-aware structural metric learning objective. In the testing stage, given a probe image, we compute the Euclidean distances between the probe and gallery images using the learned deep embeddings. Our network architecture is simpler yet effective.

In this section, we first review several widely used loss functions in Section III-A, followed by the proposed loss function in Section III-B. The optimization procedure is described in Section III-C, followed by the implementation details in Section III-D.

A. Review

In this section, we briefly review three commonly used loss functions for embedding learning: identification loss, contrastive loss, and triplet loss.

1) Identification Loss: The most popular loss function for deep embedding learning is the identification loss [67], which formulates the learning task as a multiclass classification problem. It usually corresponds an n-way softmax layer in a neural network, which yields a probability distribution over n classes and minimizes the softmax loss function, denoted by

$$\mathcal{L}(\mathbf{x}_i, y_i) = -\log \frac{e^{\mathbf{W}_{y_i}^T f(\mathbf{x}_i) + b_{y_i}}}{\sum_{j=1}^{n} e^{\mathbf{W}_j^T f(\mathbf{x}_i) + b_j}} \tag{1}$$

where y_i is the class label of sample $\mathbf{x}_i$. $f(\mathbf{x}_i)$ is the deep feature of sample $\mathbf{x}_i$ and the input of the softmax layer, $\{\mathbf{W}, \mathbf{b}\}$ is the softmax layer parameters. $\mathbf{W}_j$ denotes the jth column of the weight matrix $\mathbf{W}$ and $\mathbf{b}$ is the bias term. The number of class is n. For its simplicity, it has been applied for person re-ID in [26], [27], [29], [42], and [68]. Zheng *et al*. [68] presented a competitive baseline method for person re-ID, termed ID-discriminative embedding (IDE). However, the largest person re-ID data set at present has only less than 1000 identities in the training set. When the number of identities increases, this loss necessitates a growing number of network parameters, most of which will be discarded after training.

2) Contrastive Loss: The second option is to take paired images with a binary label $(\mathbf{x}_i, \mathbf{x}_j, y_{ij})$ as an input to formulate a contrastive loss [55]–[57] based on pairwise constraints. It aims to minimize the positive pair distance and penalize the negative pair distance that is smaller than a margin α. It is usually defined as

$$\mathcal{L}(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) = (1 - y_{ij})[\alpha - d_{ij}]_+^2 + y_{ij} d_{ij}^2 \tag{2}$$

where $y_{ij} = 1$ for a positive pair $(\mathbf{x}_i, \mathbf{x}_j)$ and $y_{ij} = 0$ for a negative pair. The operator $[\cdot]_+$ denotes the hinge loss. This loss results in an absolute distance that can answer the question "How similar/dissimilar are these two person images?" It has been investigated for person re-ID in [25], [33], and [69].

3) Triplet Loss: Triplet loss [44], [58] has attracted much more attention in recent years for its impressive performance in face recognition [44]. It can be seen as an improvement of contrastive loss, and it is measured on the triplets $\{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)\}$, where $\mathbf{x}_i$ is the *anchor* of triplet that shares the same class label with the *positive* $\mathbf{x}_j$ and has a different class label with the *negative* $\mathbf{x}_k$. It encourages the network to find an embedding space where the distance between $\mathbf{x}_i$ and $\mathbf{x}_k$ is larger than the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ by a positive margin α. The cost function that is being minimized is defined as

$$\mathcal{L}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = \left[d_{ij}^2 + \alpha - d_{ik}^2\right]_+. \tag{3}$$

Hard negative sample mining in large minibatches [44] is usually employed to select sufficient nontrivial triplets. In fact, the performance of triplet loss-based methods highly depends on the hard sample mining strategy [47]. This loss has been exploited in [37]–[40] for person re-ID. Although the triplet loss has yielded a promising performance, it usually suffers from slow convergence and poor local optima [46], partially due to the reason that it only employs three samples (one positive pair $(\mathbf{x}_i, \mathbf{x}_j)$ and one negative pair $(\mathbf{x}_i, \mathbf{x}_k)$) for distance comparison in each update.

B. Our Loss Function

In this paper, we aim to design a loss function that is able to enhance the contrastive loss and triplet loss. To address this issue, the lifted structured loss [45] is developed to improve the standard triplet loss by comparing the positive pair with all negative pairs. However, it is formulated with unnormalized deep features, yielding an unbounded Euclidean distance.

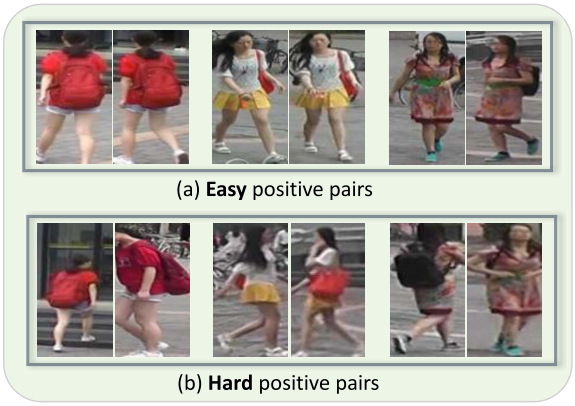

Fig. 2. Illustration of (a) easy positive pairs and (b) hard positive pairs. Images are sampled from the Market-1501 [72] data set.

We found that the lifted structured loss is sensitive to input features with large norm $\|\mathbf{x}_i\|_2$. ℓ_2 normalization can be a natural solution to avoid such situation. However, it is too stringent for the lifted structured loss [45] as well as other state-of-the-art triplet-based methods [38], [46], [70], [71], since it bounds the Euclidean distance to vary in a small range $[0, 2]$, thus making the optimization difficult and very slow.

In this paper, we introduce a structural loss that can effectively utilize the ℓ_2-normalized features for deep embedding learning. Following the merit of the lifted structured loss [45], we expect that our loss can allow each positive pair to be compared with more than one negative pairs. We use the minibatches training strategy and generate triplets online. For each iteration, given a triplet set $\mathcal{T} = \{\mathbf{x}_i, \mathbf{x}_j, \{\mathbf{x}_{ik}^-\}_{k=1}^N\}$, where $(\mathbf{x}_i, \mathbf{x}_j)$ is a positive pair and $\{\mathbf{x}_i, \mathbf{x}_{ik}^-\}_{k=1}^N$ are all the corresponding negative pairs, indexed by k, along the direction of *anchor* $\mathbf{x}_i$, we formulate the structural loss as

$$\mathcal{L}(\mathcal{T}) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \log\left(1 + \sum_{k=1}^{N} e^{\left(d_{ij}^2 - d_{ik}^2 + \alpha\right)/\xi}\right) \tag{4}$$

where α is a margin parameter. $\mathcal{P}$ is a set of positive pairs, indexed by (i, j), consisting of $|\mathcal{P}|$ positive pairs in a minibatch. $d_{ik}^2 = \|f(\mathbf{x}_i) - f(\mathbf{x}_{ik}^-)\|_2^2$ is the squared Euclidean distance. All the input features are ℓ_2-normalized. The parameter $\xi < 1$ is a scale factor introduced to yield softer triplet loss, which can accelerate the training. Without this scale factor, the optimization is more prone to collapsing unexpectedly. Similar to lifted structured loss [45], (4) is a generalization of widely used triplet loss. Its main improvement on [45] is that it can effectively utilize the ℓ_2-normalized features for training. As it employs all negative pairs for each update, we term it as ℓ_2Apair loss.

In large-scale person re-ID data sets, e.g., Market-1501 [72], each identity usually has dozens of images, manually/automatically selected from video sequences captured by multiple cameras. We found that some images of the same identity look very similar, thus resulting in many *easy* positive pairs shown in Fig. 2(a). Generally, the easy positive pairs contribute less to the training of network, since they can be recognized as the same person very easily. Existing methods

mainly focus on hard negative mining, but they pay less attention on hard positive mining. In this paper, we design a technique to pay more attention on the *hard* positive pairs for person re-ID. For a hard positive pair $(\mathbf{x}_i, \mathbf{x}_j)$ from the cth class (identity), we compare its squared Euclidean distance with a class-specific threshold to compute a hardness-aware weight β_{ij} as

$$\beta_{ij} = \exp\left(d_{ij}^2 - \tau_c\right) \tag{5}$$

where τ_c is a hard threshold of the cth class. The positive pair $(\mathbf{x}_i, \mathbf{x}_j)$ will be treated as a hard positive pair if $d_{ij}^2 > \tau_c$, which will result in larger contribution to the overall objective. The primary task is how to set the hard threshold. It is unwise to directly employ the average positive distance of class c as the threshold, since it will make nearly half of positive pairs be treated as hard and the network will be more prone to overfitting. In this paper, the hard threshold is computed as

$$\tau_c = 2\frac{1}{|\mathcal{P}_c|}\sum_{(i,j)\in\mathcal{P}_c} d_{ij}^2 - \min_{(i,j)\in\mathcal{P}_c}\left(d_{ij}^2\right) \tag{6}$$

where $|\mathcal{P}_c|$ is the set of positive pairs of the cth class. Using the hard threshold in (6), only a small proportion of positive pairs will be assigned a large weight ($\beta_{ij} > 1$), which can avoid excessive attention on hard pairs. Then, the hardness-aware ℓ_2Apair loss in (4) can be reformulated as

$$\mathcal{L}(\mathcal{T}) = \frac{1}{\mathcal{B}}\sum_{(i,j)\in\mathcal{P}} \beta_{ij}\log\left(1 + \sum_{k=1}^{N} e^{\left(d_{ij}^2 - d_{ik}^2 + \alpha\right)/\xi}\right) \tag{7}$$

where $\mathcal{B} = \sum_{(i,j)\in\mathcal{P}} \beta_{ij}$ is the sum of all positive pair weights.

The introduced positive pair weight β_{ij} in (6) enables the hard positive pair to contribute more than the easy positive pair. Although some previous works have adopted an offline hard positive samples selection strategy, easy positive samples are filtered directly, thus resulting in a waste of training samples. In this paper, all positive pairs are used, while each positive pair is adaptively assigned a hardness-aware weight to modulate its contribution.

Our proposed loss in (7) supervises the network to optimize the embedding in a local manner. It has been shown in [47] that the local manner may result in a failure that the gradient signal from the positive pair gets outweighed by the negative pairs, which leads to groups of examples with the same class label being separated into partitions in the embedding space that are far apart from each other. To alleviate this problem, we introduce a global loss term that explores the global structure of the embedding space to enhance the local loss in (7). Our global loss term takes into consideration the second-order statistics of positive/negative pair distances, defined by

$$\mathcal{L}_{global} = \frac{1}{2}\left(\left[\sigma_p^2 - \alpha_p\right]_+ + \left[\sigma_n^2 - \alpha_n\right]_+\right) \tag{8}$$

$$\sigma_p^2 = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}\left(d_{ij}^2 - \mu_p\right)^2 \tag{9}$$

$$\sigma_n^2 = \frac{1}{|\mathcal{N}|}\sum_{(i,l)\in\mathcal{N}}\left(d_{il}^2 - \mu_n\right)^2 \tag{10}$$

where $|\mathcal{N}|$ denotes the number of all negative pairs in the negative pair set $\mathcal{N}$, and σ_p^2 and σ_n^2 denote the variances of positive pair distances and negative pair distances, respectively. μ_p and μ_n denote the mean value of positive pair distance and negative pair distance, respectively, in the normalized Euclidean distance space. α_p and α_n are two margin parameters. The values of μ_p and μ_n are hard to be estimated. It is too stringent to set them as fixed values. In this paper, the values of μ_p and μ_n are updated for each iteration as

$$\mu_p^t = \gamma\,\mu_p^{t-1} + (1-\gamma)\mu_p^t \tag{11}$$

$$\mu_u^t = \gamma\,\mu_u^{t-1} + (1-\gamma)\mu_u^t \tag{12}$$

where μ_p^t (μ_u^t) denotes the mean value of positive (negative) pair distances estimated using the tth minibatch. γ is a parameter restricted in $[0, 1]$ that controls the updating of μ_p^t and μ_u^t.

The introduced global loss term in (8) is designed to penalize the large variance of positive/negative pair distances in the normalized Euclidean distance space. This global loss term is able to regularize the network and improve its generalization capability by combining the global loss term in (8) and the local loss term in (4).

The final objective function is formulated by

$$\mathcal{L}(\mathcal{T}) = \frac{1}{\mathcal{B}}\sum_{(i,j)\in\mathcal{P}} \beta_{ij}\log\left(1 + \sum_{k=1}^{N} e^{\left(d_{ij}^2 - d_{ik}^2 + \alpha\right)/\xi}\right) + \frac{\lambda}{2}\left(\left[\sigma_p^2 - \alpha_p\right]_+ + \left[\sigma_n^2 - \alpha_n\right]_+\right) \tag{13}$$

where the parameter λ is used to balance the two terms. Note that the global loss term is not a hard constraint. If λ is too large, the signal of gradient from the second term will outweigh the signal of the first term, which makes the network prone to overfitting.

The mathematical formulation of our final objective in (13) is a generalization of triplet loss. It is designed with ℓ_2 normalized features and allows the positive pair to be compared against all the negative pairs. It adaptively assigns larger weights to hard positive pairs, which facilitates the training. It also includes a global loss term that controls the variance of positive/negative pair distances, thus improving the generalization capability of the network. By this way, the learning performance can be improved effectively.

C. Optimization

To solve the optimization problem in (13), we apply the stochastic gradient descent scheme to update the gradient of objective $\mathcal{L}$ with respect to the feature embedding $f(\cdot)$. We summarize the backpropagation procedure in Algorithm 1, where $\mathcal{F}_{ij} = \log(1 + \sum_{k=1}^{N} e^{(d_{ij}^2 - d_{ik}^2 + \alpha)/\xi})$ for simple expression, and

$$\frac{\partial\mathcal{L}}{\partial\sigma_p^2} = \frac{\lambda}{2}\mathbb{1}\left[\sigma_p^2 > \alpha_p\right] \tag{14}$$

$$\frac{\partial\mathcal{L}}{\partial\sigma_n^2} = \frac{\lambda}{2}\mathbb{1}\left[\sigma_n^2 > \alpha_n\right] \tag{15}$$

Algorithm 1 Backpropagation

Input: A mini-batch $\{\mathbf{x}_i\}_{i=1}^m$ with multiclass labels; margin α; parameter λ; pair weights $\{\beta_{ij}\}$; dense pairwise squared distance matrix $D = \{d_{ij}^2\}$;

Output: The gradients $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_i)}$, $\forall i \in [1, m]$.

Initialization: $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_i)} = 0$, $\forall i \in [1, m]$;

for $i = 1, \cdots, m$ **do**

  for $j = 1, \cdots, m$, *s.t.* $(i, j) \in \mathcal{P}$ **do**

    $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_i)} \leftarrow \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_i)} + \frac{\lambda}{2}\frac{\partial \mathcal{L}}{\partial \sigma_p^2}\frac{\partial \sigma_p^2}{\partial d_{ij}^2}\frac{\partial d_{ij}^2}{\partial f(\mathbf{x}_i)}$;

    $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_j)} \leftarrow \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_j)} + \frac{\lambda}{2}\frac{\partial \mathcal{L}}{\partial \sigma_p^2}\frac{\partial \sigma_p^2}{\partial d_{ij}^2}\frac{\partial d_{ij}^2}{\partial f(\mathbf{x}_j)}$;

    for $k = 1, \cdots, m$, *s.t.* $(i, k) \in \mathcal{N}$ **do**

      $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_i)} \leftarrow \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_i)} + \frac{\beta_{ij}}{\mathcal{B}}\left(\frac{\partial \mathcal{F}_{ij}}{\partial d_{ij}^2}\frac{\partial d_{ij}^2}{\partial f(\mathbf{x}_i)} + \frac{\partial \mathcal{F}_{ij}}{\partial d_{ik}^2}\frac{\partial d_{ik}^2}{\partial f(\mathbf{x}_i)}\right)$;

      $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_j)} \leftarrow \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_j)} + \frac{\beta_{ij}}{\mathcal{B}}\frac{\partial \mathcal{F}_{ij}}{\partial d_{ij}^2}\frac{\partial d_{ij}^2}{\partial f(\mathbf{x}_j)}$;

      $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_k)} \leftarrow \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_k)} + \frac{\beta_{ij}}{\mathcal{B}}\frac{\partial \mathcal{F}_{ij}}{\partial d_{ik}^2}\frac{\partial d_{ik}^2}{\partial f(\mathbf{x}_k)}$;

  for $l = 1, \cdots, m$, *s.t.* $(i, l) \in \mathcal{N}$ **do**

    $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_i)} \leftarrow \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_i)} + \frac{\lambda}{2}\frac{\partial \mathcal{L}}{\partial \sigma_n^2}\frac{\partial \sigma_n^2}{\partial d_{il}^2}\frac{\partial d_{il}^2}{\partial f(\mathbf{x}_i)}$;

    $\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_l)} \leftarrow \frac{\partial \mathcal{L}}{\partial f(\mathbf{x}_l)} + \frac{\lambda}{2}\frac{\partial \mathcal{L}}{\partial \sigma_n^2}\frac{\partial \sigma_n^2}{\partial d_{il}^2}\frac{\partial d_{il}^2}{\partial f(\mathbf{x}_l)}$;

$$\frac{\partial \sigma_p^2}{\partial d_{ij}^2} = \frac{2}{|\mathcal{P}|}\left(d_{ij}^2 - \mu_p\right) \tag{16}$$

$$\frac{\partial \sigma_n^2}{\partial d_{il}^2} = \frac{2}{|\mathcal{N}|}\left(d_{il}^2 - \mu_n\right) \tag{17}$$

$$\frac{\partial \mathcal{F}_{ij}}{\partial d_{ij}^2} = \frac{e^{\mathcal{F}_{ij}} - 1}{\xi e^{\mathcal{F}_{ij}}} \tag{18}$$

$$\frac{\partial \mathcal{F}_{ij}}{\partial d_{ik}^2} = -\frac{e^{\left(d_{ij}^2 - d_{ik}^2 + \alpha\right)/\xi}}{\xi e^{\mathcal{F}_{ij}}} \tag{19}$$

where $\mathbb{1}[\cdot]$ is an indicator function which outputs 1 if the expression evaluates to true and outputs 0 otherwise. The gradients of positive pair distance with respect to the feature embeddings are computed as

$$\frac{\partial d_{ij}^2}{\partial f(\mathbf{x}_i)} = 2(f(\mathbf{x}_i) - f(\mathbf{x}_j)) \tag{20}$$

$$\frac{\partial d_{ij}^2}{\partial f(\mathbf{x}_j)} = 2(f(\mathbf{x}_j) - f(\mathbf{x}_i)). \tag{21}$$

D. Implementation Details

1) Network Architecture: We use a subnetwork of GoogLeNet architecture (*Inception v1* [73]), from the image input to the output of *inception-4e*, followed by a global average pooling layer, a fully connected layer, an ℓ_2-normalization layer, and finally the loss layer. Specifically, the person images are resized to 160×80 as an input. In the testing stage, the output of the ℓ_2-normalization layer is directly used to compute the distance for person retrieval. For data preprocessing, we use the standard horizontal flips of the resized images.

2) Network Training: We use the Caffe [74] package for the implementation. We pretrain the network on the ImageNet ILSVRC data set and fine-tune it on the person re-ID data sets. In this paper, we adopt an online sampling strategy for minibatch generation. For each sequentially selected identity, we randomly select at most K images into a minibatch. When all identities have been traversed, we shuffle the identity list for next round. The batch size is 150 and each identity has at most five images in a minibatch. Besides, the triplet set is also generated online. We set the initial learning rate as 0.01 and divide it by 10 after 10 000 iterations and 12 500 iterations. The weight decay is 0.0002 and the momentum for gradient update is 0.9. Each model is trained for 15 000 iterations. The margin parameter α, parameter λ, and scale factor ξ are set to 0.2, 0.5, and 0.05, respectively. The parameter γ is set to 0.95 for stable updating of mean distance. The parameters α_p and α_n are set to 0.01 and 0.1, respectively. The final embedding dimension of our approach is 128, which facilitates person retrieval on large-scale person re-ID data sets.

IV. Experiments

A. Data Sets and Evaluation Protocols

The evaluation is carried out on three large-scale person re-ID data sets: Market-1501 [72], DukeMTMC-reID [26], and CUHK03 [32].

Market-1501 [72] is one of the largest person re-ID data sets, containing 32 668 bounding boxes (cropped images) of 1501 identities. All the bounding boxes are detected by the deformable part model (DPM) pedestrian detector [75]. Each identity has multiple images captured by at least two cameras and at most six cameras. Specifically, the training set contains 12 936 bounding boxes of 750 identities. The testing set contains 19 732 bounding boxes of 751 identities, where only one image of each identity is randomly selected as a query image for each camera. In total, the testing set contains 3368 query images. There are 2793 images included as distractors in the original gallery set for testing.

DukeMTMC-reID [26] is a new large-scale person re-ID data set, derived from a multicamera pedestrian tracking data set (DukeMTMC [76]). It contains 36 411 hand-drawn bounding boxes of 1812 identities, taken from eight different camera views, in which 1404 identities appear in more than two camera views and 408 identities appear in only one camera view whose images are used as distractors; 16 522 bounding boxes of 702 identities are randomly selected for training and the rest are used for testing, including 2228 query images and 17 661 gallery images.

CUHK03 [32] is a widely used re-ID data set, containing 13 164 bounding boxes of 1360 identities, captured by two disjoint cameras. In the original evaluation protocol [32], 1160 identities are employed for training, and the remaining identities are used for validation and testing. In this paper, we adopt a new data set split [20] for CUHK03, where 767 identities are used for training and the remaining 700 identities are employed for testing. The new protocol is more

(a) (b) (c)

Fig. 3. Sample images from three large-scale person re-ID data sets. (a) Market-1501 [72]. (b) DukeMTMC-reID [26]. (c) CUHK03 [32].

TABLE I

TOP-RANKED AVERAGE RECOGNITION RATES (CMC@RANK-R, %) AND MAP (%) OF ℓ_2APAIR LOSS AND ITS TWO VARIANTS ON THE MARKET-1501, DUKEMTMC-REID, AND CUHK03 DATA SETS. A LARGER NUMBER INDICATES A BETTER RESULT

Iterations	Methods	Market-1501				DukeMTMC-reID				CUHK03			
		R=1	R=5	R=10	mAP	R=1	R=5	R=10	mAP	R=1	R=5	R=10	mAP
10000	ℓ_2Apair	77.76	90.53	93.82	58.93	68.13	81.46	85.77	49.04	39.96	61.83	70.91	37.97
	ℓ_2Apair+HA	79.42	90.97	94.12	60.12	70.02	83.26	87.70	50.04	40.31	63.33	72.69	39.11
	ℓ_2Apair+HA+GL	79.57	91.03	94.18	60.68	71.01	84.61	88.29	51.87	41.10	64.26	73.12	39.64
15000	ℓ_2Apair	79.01	90.94	94.03	61.10	71.18	83.44	87.88	51.56	40.53	63.05	71.34	38.69
	ℓ_2Apair+HA	79.84	91.03	94.30	61.45	71.90	84.34	88.51	52.10	41.32	63.12	73.20	39.65
	ℓ_2Apair+HA+GL	80.73	91.89	94.66	62.57	72.53	85.23	88.29	53.18	42.32	65.33	73.77	40.06

challenging, since the number of training images becomes much less.

Fig. 3 shows some sample images from these data sets. We use the widely used evaluation protocol [35]. In the matching process, we calculate the similarities between each query and all the gallery images and return the rank list according to the similarities. All the experiments are under the single query setting. The performance is evaluated by the cumulative matching characteristics (CMC) curve that is an estimation of the expectation of finding the correct match in the top-K matches and the mean average precision (mAP) score. We use the evaluation codes provided by Zhao *et al.* [77] for the empirical evaluation in Section IV-B.

B. Empirical Analysis

1) Effect of the Hardness-Aware Weights: In Section III-B, we present a hard positive mining strategy that aims to pay more attention on hard positive pairs during optimization. As shown in (5), for each positive pair, we compare its squared Euclidean distance d_{ij}^2 with a class-specific hard threshold τ_c. Then, the positive pair with a large distance ($d_{ij}^2 > \tau_c$) will result in larger contribution to the overall objective. In this section, we evaluate its effect by ablation study on the three benchmark data sets. The learning objective in (7) that has the hardness-aware weight β_{ij} is termed ℓ_2Apair + HA. In Table I, we compare its performance with the ℓ_2Apair loss in (4). Note that we compare their performance using the networks trained with 10 000 iterations and 15 000 iterations.

We can observe in Table I that, by 10 000 iterations, ℓ_2Apair + HA improves ℓ_2Apair by 1.7%, 1.9%, and 0.4% rank-1, and 1.2%, 1%, and 1.2% mAP score on the three data sets, respectively. By 15 000 iterations, ℓ_2Apair + HA improves ℓ_2Apair by 0.8%, 0.7%, and 0.8% rank-1 and 0.4%, 0.5%, and 1% mAP. It shows that the effect of the proposed hard positive mining technique is more obvious in the early optimization stage. By exploiting this hardness-aware weighting technique, *hard* positive points can be pulled closer to the anchor points, thus accelerating the optimization, especially in the early learning stage. After 10 000 iterations, the intraclass distance has been significantly reduced. The margin between the hard positive pairs and the easy positive pairs becomes much smaller. Then, the penalization on the hard positive pair becomes weak. It can avoid excessive focus on the hard samples.

2) Effect of the Global Loss Term: This paper presents a global loss term that aims to regularize the network by utilizing the second-order statistics of positive/negative distance distribution. In detail, we expect the the learned distance has a low variance in the final embedding space. It is mainly inspired by the observation in [47] that local learning manner may result in a failure: the gradient signal from positive pairs may get outweighed by negative pairs. It will lead to groups of examples with the same class label being separated into partitions in the embedding space.

By the proposed regularization term, we observe in Table I that the performance of ℓ_2Apair + HA can

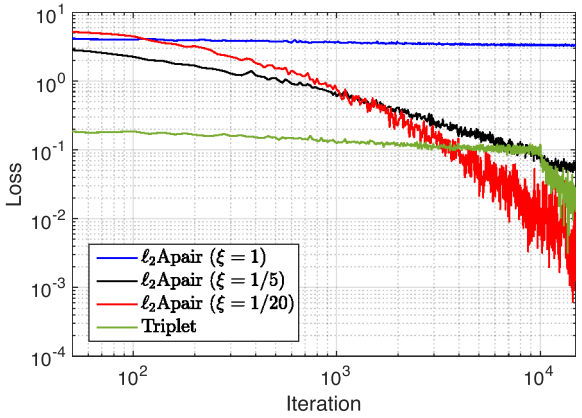

Fig. 4. Convergence curves of ℓ_2Apair loss with $\xi = 1$, 1/5, and 1/20, and triplet loss.

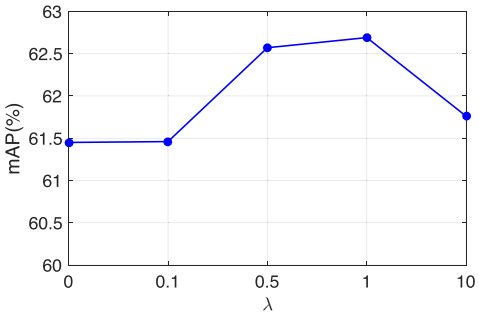

Fig. 5. MAP comparison at varying $\lambda = \{0, 0.1, 0.5, 1, 10\}$.

TABLE II
PERFORMANCE COMPARISON (CMC@RANK-R AND MAP, %) ON THE MARKET-1501 DATA SET AT VARYING EMBEDDING DIMENSIONS

Dimension	32	64	128	256	512	1024
Rank-1	73.96	76.84	80.73	81.71	82.90	82.33
mAP	56.47	58.54	62.57	63.51	64.63	64.74

be further improved. By 15 000 iterations, rank-1 is improved by 0.9%, 0.6%, and 1% and mAP is improved by 1.1%, 1%, and 0.4%, respectively, on the three data sets. The improvement is actually not trivial, since ℓ_2Apair loss is already a strong baseline which exploits the structural relationship in the data space. It is indicated in Section IV-C that our method has outperformed most state-of-the-art methods while using much lower dimensional features and a less sophisticated base network structure.

3) Effect of the Scale Factor ξ: In (4), the ℓ_2Apair loss introduces a scale factor ξ to make the triplet loss softer in the ℓ_2-normalized distance space, which can avoid the optimization to be collapsing. To evaluate its impact on the convergence, we plot the training loss curves with $\xi = 1$, 1/5, and 1/20, respectively, in Fig. 4. The loss curve of triplet loss is also shown in Fig. 4. We can see that without this scale factor ($\xi = 1$), the training loss drops very slow due to the collapsing of optimization. When we set $\xi = 1/5$, the loss drops fast since the loss has been fivefolds amplified. In the experiment, we set $\xi = 1/20$, which shows a faster convergence speed than $\xi = 1/5$. It is recommended to set this scale factor as 1/10 or 1/20 for a faster optimization. Too small value of ξ will make the output of exp(·) beyond the limit of precision of double/float type, thus making the optimization unstable. Besides, we also observe in Fig. 4 that ℓ_2Apair loss ($\xi = 1/20$) converges much faster than standard triplet loss, mainly benefiting from the structural distance comparison and the soft-margin design.

4) Effect of the Parameter λ: The parameter λ in (13) makes a tradeoff between the ℓ_2Apair loss term and the proposed global regularization term. In this section, we investigate its effect by comparing the performance of our loss [see (13)] at varying $\lambda = \{0, 0.1, 0.5, 1, 10\}$. As shown in Fig. 5, we can observe a clear improvement when $\lambda = 0.5$ or $\lambda = 1$. When λ is small, its effect is not significant. A too large λ will also make the gradient signal from the second term (global loss) outweigh the first term (local loss), which makes the optimization slow.

5) On the Embedding Dimension d: In Table II, we evaluate the performance of our approach [based on (13)] at varying embedding sizes {32, 64, 128, 256, 512, 1024} on the Market-1501 data set. As shown in Table II, our approach has the ability of learning much lower dimensional discriminative embeddings. Our approach can obtain a very impressive performance: 73.96% rank-1 and 56.47% mAP, even when $d = 32$. When $32 \leq d \leq 128$, the performance increases significantly. When $128 \leq d \leq 1024$, the performance increases much slower. Rank-1 starts to drop when $d > 512$. A larger embedding size may not guarantee a better performance, but it undoubtedly incurs a higher time complexity in the testing stage. Therefore, we set the final embedding size as 128, which works very well in the experiments. Such a low-dimensional embedding facilitates the large-scale person retrieval in the real-world scenarios.

C. Comparison With the State of the Arts

In this section, we compare our method with the recent state-of-the-art results on the Market-1501, DukeMTMC-reID, and CUHK03 data sets. Note that, for a fair comparison, we train the network using the same input image size (224×224) with that of most state-of-the-art methods, and use the standard evaluation codes of [72]. Experimental results are presented in Tables III–V.

1) Results on Market-1501: In Table III, we compare our result with state-of-the-art methods on the Market-1501 data set. Note that, all the state-of-the-art methods in Table III are based on deep learning techniques. As shown in Table III, the competitive methods, such as IDE [68], SVDNet [80], and PAN [81], are font of using identification loss together with the ResNet architecture. IDE outputs a strong baseline result: 73.8% rank-1 and 47.9% mAP. The SVDNeT incorporates singular vector decomposition into IDE to orthogonalize the base components in the fully connected layer. It significantly improves the performance of IDE, but it also largely increases the computational complexity. The PAN [81] method introduces the pedestrian alignment into the IDE model. It can

TABLE III

PERFORMANCE COMPARISON (CMC@RANK-R AND MAP, %) ON THE MARKET-1501 DATA SET. A LARGER NUMBER INDICATES A BETTER RESULT. † DENOTES THE RESULTS FROM CONCURRENT WORK ONLY PREPUBLISHED ON ARXIV. * DENOTES RESULTS USING THE EVALUATION CODES OF [77]. I, V, C, AND T STAND FOR IDENTIFICATION, VERIFICATION, CONTRASTIVE, AND TRIPLET, RESPECTIVELY. DIM MEANS THE DIMENSION OF THE LEARNED DEEP FEATURES

Market 1501	Dim	R=1	R=5	R=10	mAP
AttentionNet(T) [34]	512	48.24	-	-	24.43
MSTriplet(T) [39]	500	45.10	70.10	78.40	-
DeHist [78]	512	59.47	80.73	86.94	-
Contrastive(C) [33]	-	62.32	-	-	36.23
GatedSiamese(C) [33]	-	65.88	-	-	39.55
IDE(ResNet50)(I) [68]	2048	73.80	87.60	91.30	47.90
Dem(GoogleNet)(I+V) [79]	1024	73.52	-	-	51.15
Dem(ResNet50)(I+V) [79]	4096	79.51	-	-	59.87
SVDNet(ResNet50)(I) [80]	2048	82.30	92.3	95.20	62.10
Triplet*(GoogleNet)(T) [77]	512	75.90	89.30	92.90	54.30
PartAN*(GoogleNet)(T) [77]	512	81.00	92.00	94.70	63.40
PAN(ResNet50)(I)† [81]	2048	82.81	93.53	-	63.35
MSML(ResNet50)(T)† [71]	1024	**85.20**	**93.70**	-	**69.60**
Ours-Eq. (13)*(GoogleNet)	128	84.26	93.59	**95.99**	67.31
Ours-Eq. (13)(GoogleNet)	128	83.43	93.14	95.37	63.66

reduce the negative effect of the background. It obtains 82.81% rank-1 and 63.4% mAP. Although they yield a promising performance on Market-1501, they rely on the ResNet50 architecture that is deeper and complicated than GoogleNet. ResNet-50 has a much higher memory and computation requirements than GoogleNet, especially while training. The size of the final trained model becomes an important concern to consider if we are looking to deploy a model to run locally on mobile. Besides, they usually use the output of the last pooling layer as the deep feature for testing, thus resulting in a 2048-D embedding which is unsuitable for fast person retrieval. However, this paper aims to learn low-dimensional deep embeddings using a relatively small network architecture: GoogleNet(V1). As shown in Table III, our method obtains 83.43% rank-1 and 63.66% mAP using only 128-D features. It has surpassed most state-of-the-art results listed in Table III. Xiao *et al*. [71] presented a generalized triplet loss with extremely hard sample mining. It obtains a better performance than our method. However, it is implemented using ResNet50 and outputs 1024-D features. The Dem [79] method employs the complementation of identification loss and verification loss and also reports its result using GoogleNet: 73.52% rank-1 and 51.15% mAP. Our method surpasses it by nearly 10% rank-1 and over 12% mAP. Zhao *et al*. [77] designed a part-aligned network module for pedestrian feature learning based on GoogleNet and triplet loss. Without the part detector, its base network is the same as ours. We improve its baseline (GoogleNet+Triplet loss) by over 8% rank-1 and 13% mAP under the same evaluation protocol. Even with the part-detector module, its performance is still inferior to ours.

The competitive performance on the Market-1501 data set has demonstrated the effectiveness of the proposed approach. It can be mainly attributed to the proposed hardness-aware structural learning objective and the global loss term.

Fig. 6. Barnes–Hut t-SNE visualization [82] of our learned embeddings on a subset of the testing set in the Market-1501 data set [72]. Best viewed in color.

TABLE IV

PERFORMANCE COMPARISON (CMC@RANK-R AND MAP, %) ON THE DUKEMTMC-REID DATA SET. A LARGER NUMBER INDICATES A BETTER RESULT. † DENOTES RESULTS FROM CONCURRENT WORK ONLY PREPUBLISHED ON ARXIV. I STANDS FOR IDENTIFICATION

DukeMTMC-ReID	Dim	R=1	R=5	R=10	mAP
IDE(CaffeNet)(I)† [68]	4096	46.90	63.20	69.20	28.30
IDE(ResNet50)(I)† [68]	2048	65.50	78.50	82.50	44.10
OIM(GoogleNet)(I) [29]	256	61.70	-	-	-
OIM(ResNet50)(I) [29]	256	68.10	-	-	-
GAN(ResNet50)(I) [26]	2048	67.68	-	-	47.13
PAN(ResNet50)(I)† [81]	2048	71.59	83.89	-	51.51
SVDNet(ResNet50)(I) [80]	2048	**76.70**	86.40	89.90	**56.80**
Ours-Eq. (13)(GoogleNet)	128	75.40	**87.66**	**90.98**	54.76

Fig. 6 shows the Barnes–Hut t-distributed stochastic neighbor embedding (t-SNE) visualization [82] on a subset of the testing set using the learned deep features. It is clearly shown that most visually similar person images are grouped together. Our method is able to learn semantically meaningful features.

2) Results on DukeMTMC-reID: The DukeMTMC-reID data set is a newly released large-scale person re-ID data set, derived from a multicamera pedestrian tracking data set (DukeMTMC [76]). We compare our method with several state-of-the-art DNN-based methods on this data set in Table IV. All the compared methods listed in Table IV are based on the identification loss. The IDE model [68] reports its results using two network architectures: 46.9% rank-1 and 28.3% mAP based on CaffeNet, and 65.5% rank-1 and 44% mAP based on ResNet50. It shows that as a state-of-the-art network architecture, ResNet50 can bring a significant performance improvement on the simple CNN architecture, such as CaffeNet. Another similar case is the Online Instance Matching (OIM) [29] model which obtains 61.7% rank-1 with GoogleNet and 68.1% rank-1 with ResNet50. Nearly, 7% improvement is observed. The other three methods all adopt ResNet50 as their base network, in which Zheng *et al*. [26] utilized the generative adversarial network technique to generate a large amount of unlabeled pedestrian images. These unlabeled images are further employed to augment the training

TABLE V

PERFORMANCE COMPARISON (CMC@RANK-R AND MAP, %) ON THECUHK03 (DETECTED) DATA SET USING THE NEW EVALUATION PROTOCOL [20]. THIS PROTOCOL USES A LARGER TESTING SET (767 PERSONS FOR TRAINING AND 700 PERSONS FOR TESTING). A LARGER NUMBER INDICATES A BETTER RESULT. † DENOTES RESULTS FROM CONCURRENT WORK ONLY PREPUBLISHED ON ARXIV. I STANDS FOR IDENTIFICATION

CUHK03	Dim	R=1	mAP
IDE(CaffeNet)(I) [68]	4096	15.60	14.90
IDE(ResNet50)(I) [68]	2048	30.50	21.10
PAN(ResNet50)(I)† [81]	2048	36.30	34.00
SVDNet(CaffeNet)(I) [80]	4096	27.70	24.90
SVDNet(ResNet50)(I) [80]	2048	**41.50**	**37.30**
Ours-Eq. (13)(GoogleNet)	128	39.64	35.67

set of the IDE model, which yields 2% performance improvement. Our method adopts a much smaller network architecture: GoogleNet, but achieves a state-of-the-art result on DukeMTMC-reID. Our rank-1 accuracy is 75.4% that surpasses the strong baseline [IDE (ResNet50)] by nearly 10%. It experimentally demonstrates the effectiveness of our loss. The main reason is that the identification loss aims to learn a nonlinear mapping from person image to person ID and does not consider the interpersonal relationship directly. That is, the value of training data has not been fully exploited. Our loss aims to optimize a relative distance by comparing the positive pair with all the corresponding negative pairs. The structural relationship among samples has been exploited effectively.

3) Results on CUHK03: In this paper, we apply a new testing protocol in [20] for the evaluation on the CUHK03 data set. This new protocol uses a smaller training set and a larger testing set (767 persons for training and 700 persons for testing). It is more challenging than the original setting [32] where 1367 persons are employed for training and only 100 persons are used for testing. This data set includes two subsets in which one is manually labeled and the other one is detected by the DPM pedestrian detector. We only report our result on the detected set, since we observe a similar performance on the two sets. Under the new data set split, only 7365 images can be used for training, thus making the network prone to overfitting. In each iteration, we randomly sample a minibatch and generate large numbers of triplets online for our learning objective, which can reduce the risk of overfitting. It can be observed from the performance comparison in Table V, the IDE model [68] can only obtain 30.50% rank-1 and 21.1% mAP using ResNet50, while our method achieves 39.64% rank-1 and 35.67% mAP using GoogleNet. It improves the mAP of IDE by over 14%. It can be mainly attributed to the structural learning objective and the global loss term in this paper. A low-dimensional but discriminative deep embedding can be learned by our method. In Table V, the SVDNet [80] obtains the best result on this data set by orthogonalizing the base components of fully connected layer. However, if taking the network structure and feature dimension into consideration, our method is more competitive than SVDNet.

V. CONCLUSION

In this paper, we present an effective person re-ID framework by discriminatively learning a nonlinear deep feature mapping from person images to low-dimensional embeddings, where similar samples are mapped closer to each other, while dissimilar samples are pushed farther apart. The proposed approach jointly learns feature representation and distance metric in an end-to-end manner. The main contribution of this paper is that we develop a hardness-aware structural metric learning objective where each positive pair is allowed to be compared with all the corresponding negative pairs within minibatch and each positive pair is assigned a hardness-aware weight to adaptively modulate its contribution. Moreover, we incorporate a global loss term that penalizes large variance of positive/negative pair distances into the proposed objective function, which improves the generalization capability of the network effectively. Extensive experimental evaluations on three large-scale data sets have demonstrated the effectiveness of our approach.

REFERENCES

[1] L. An, X. Chen, S. Yang, and X. Li, "Person re-identification by multi-hypergraph fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2763–2774, Nov. 2017.

[2] X. Li, L. Liu, and X. Lu, "Person reidentification based on elastic projections," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1314–1327, Apr. 2018.

[3] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1288–1296.

[4] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[5] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1089–1102, Jun. 2017.

[6] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1363–1372.

[7] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.

[8] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.

[9] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 536–551.

[10] R. R. Varior, G. Wang, J. Lu, and T. Liu, "Learning invariant color features for person reidentification," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3395–3410, Jul. 2016.

[11] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3685–3693.

[12] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.

[13] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.

[14] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.

[15] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3610–3617.

[16] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.

[17] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1268–1277.

[18] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 144–151.

[19] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1741–1750.

[20] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[21] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2666–2672.

[22] Q. Wang, W. Zuo, L. Zhang, and P. Li, "Shrinkage expansion adaptive metric learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 456–471.

[23] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

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

[25] H. Shi *et al.*, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 732–748.

[26] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 3754–3762.

[27] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3346–3355.

[28] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1249–1258.

[29] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3376–3385.

[30] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2194–2200.

[31] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. IEEE 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 34–39.

[32] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[33] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 791–808.

[34] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[35] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3908–3916.

[36] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2667–2675.

[37] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[38] A. Hermans, L. Beyer, and B. Leibe. (2017). "In defense of the triplet loss for person re-identification." [Online]. Available: https://arxiv.org/abs/1703.07737

[39] J. Liu *et al.*, "Multi-scale triplet CNN for person re-identification," in *Proc. ACM Multimedia Conf.*, Oct. 2016, pp. 192–196.

[40] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1320–1329.

[41] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, 2018, Art. no. 13.

[42] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. (2017). "Improving person re-identification by attribute and identity learning." [Online]. Available: https://arxiv.org/abs/1703.07220

[43] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[44] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[45] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4004–4012.

[46] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1849–1857.

[47] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy, "Deep metric learning via facility location," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2206–2214.

[48] S. Ying, Z. Wen, J. Shi, Y. Peng, J. Peng, and H. Qiao, "Manifold preserving: An intrinsic approach for semisupervised distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2731–2742, Jul. 2018.

[49] Y. Wang, W. Zhang, L. Wu, X. Lin, and X. Zhao, "Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 57–70, Jan. 2017.

[50] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[51] F. Wang, W. Zuo, L. Zhang, D. Meng, and D. Zhang, "A kernel classification framework for metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 1950–1962, Sep. 2015.

[52] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 945–954.

[53] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2808–2817.

[54] B. Shi, X. Bai, W. Liu, and J. Wang, "Face alignment with deep regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 183–194, Jan. 2016.

[55] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.

[56] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.

[57] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 539–546.

[58] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[59] O. Rippel, M. Paluri, P. Dollar, and L. Bourdev, "Metric learning with adaptive density discrimination," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015.

[60] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 791–805, Feb. 2018.

[61] X. Yang, M. Wang, R. Hong, Q. Tian, and Y. Rui, "Enhancing person re-identification in a self-trained subspace," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 3, p. 27, 2017.

[62] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[63] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.

[64] M. Geng, Y. Wang, T. Xiang, and Y. Tian. (2016). "Deep transfer learning for person re-identification." [Online]. Available: https://arxiv.org/abs/1611.05244

[65] L. Wu, C. Shen, and A. V. D. Hengel. (2016). "Personnet: Person re-identification with deep convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1601.07255

[66] H. Jin, X. Wang, S. Liao, and S. Z. Li. (2017). "Deep person re-identification with improved embedding and efficient training." [Online]. Available: https://arxiv.org/abs/1705.03332

[67] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.
[68] L. Zheng, Y. Yang, and A. G. Hauptmann. (2016). "Person re-identification: Past, present and future." [Online]. Available: https://arxiv.org/abs/1610.02984
[69] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 135–153.
[70] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2612–2620.
[71] Q. Xiao, H. Luo, and C. Zhang. (2017). "Margin sample mining loss: A deep learning based method for person re-identification." [Online]. Available: https://arxiv.org/abs/1710.00478
[72] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.
[73] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[74] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[75] H. Cho, P. E. Rybski, A. Bar-Hillel, and W. Zhang, "Real-time pedestrian detection with deformable part models," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2012, pp. 1035–1042.
[76] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 17–35.
[77] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 3219–3228.
[78] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4170–4178.
[79] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 13:1–13:20, 2017.
[80] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3820–3828.
[81] Z. Zheng, L. Zheng, and Y. Yang. (2017). "Pedestrian alignment network for large-scale person re-identification." [Online]. Available: https://arxiv.org/abs/1707.00408
[82] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.

Xun Yang is currently pursuing the Ph.D. degree with the School of Computer and Information Engineering, Hefei University of Technology, Hefei, China.

From 2015 to 2017, he was a Visiting Research Student with the Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo NSW, Australia. His current research interests include person reidentification, multimedia content analysis, computer vision, and pattern recognition.

Peicheng Zhou received the B.S. degree from the Xi'an University of Technology, Xi'an, China, in 2011, and the M.S. degree from Northwestern Polytechnical University, Xi'an, in 2014, where he is currently pursuing the Ph.D. degree.

His current research interests include computer vision and pattern recognition.

Meng Wang (SM'17) received the B.E. and Ph.D. degrees from the Special Class for the Gifted Young, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively.

He is currently a Professor with the Hefei University of Technology, Hefei. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored over 200 book chapters and journal and conference papers in these areas.

Dr. Wang was a recipient of the ACM SIGMM Rising Star Award 2014. He is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and the IEEE TRANSACTIONS ON MULTIMEDIA.