IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 28, 2019, accepted December 22, 2019, date of publication December 26, 2019,
date of current version January 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962581

Deep Multi-Task Transfer Network for Cross Domain Person Re-Identification

HUAN WANG[1] **AND JINGBO HU**[2]
[1]School of Computer Science and Technology, Baoji University of Arts and Sciences, Baoji 721016, China
[2]School of Electronic and Electrical Engineering, Baoji University of Arts and Sciences, Baoji 721016, China

Corresponding author: Jingbo Hu (jingbo_hu@126.com)

This work was supported in part by the Projects of Shaanxi Science and Technology Commission under Grant 2017JQ6030 and Grant 2018GY-050, in part by the Projects of Shaanxi Education Commission under Grant 16JS008, in part by the Projects of the Science and Technology Research and Development Program of Baoji City under Grant 2018-JH12, in part by the Key Project of Baoji University of Arts and Sciences under Grant ZK16053, in part by the Projects of Shuangbai and Industry-University-Research under Grant 209011024, and in part by the Projects of industry-university-research innovation fund of Ministry of education science and technology development center under Grant 2018A02012.

ABSTRACT As a prominent application of surveillance video analysis, person re-identification attracts much more research attention recently. Existing person re-identification models often focus on supervision by the pedestrian identity annotation, while it has limited scalability in realistic. Though several unsupervised person re-identification researches pay attention to solve this problem, they are either clustering based or cross domain based approaches, where a conventional assumption of them is the identity number of the target dataset is acknowledged. To relax this hypothesis, we propose a Deep Multi-task Transfer Network (DMTNet) for cross domain person re-identification, which conduct classification, attribute attention and identification task between source and target domains. There are three main novelties in DMTNet, including clustering number estimating algorithm to learn prior knowledge from source data to estimate the identity number, attribute attention importance learning rather than directly utilizing attribute information, and a multi-task transfer learning mechanism to transfer specific tasks cross domains. To prove the superiority of our DMTNet, we implement several compared experiments on DukeMTMC-reID and Market-1501 datasets, which results show the advancement of our network. Moreover, the discussions for different modules also point out the significance of the specific tasks.

INDEX TERMS Cross domain person re-identification, multi-task transfer, attribute attention, identity number estimating.

I. INTRODUCTION

Person re-identification (re-id) is an extremely prominent technology in surveillance system due to its significance in pedestrian retrieval, such as criminal tracking, pedestrian locating. The goal of person re-id is to identity the specific identity in gallery images, given a single probe image. This task is confronted with formidable challenges on account of severe variations in resolution, view-point, pose, occlusion and illumination across different cameras, on which most person re-id models focus to solve. Figure 1 illustrates the challenges in different datasets. Though existing person re-id approaches achieve an adequate performance, they need large amounts of annotations to train the models, which are under supervised framework. That seriously confines the scalability of person re-id in realistic scene.

To expand the promptness of the person re-id models, a number of researches pay attention to unsupervised person re-identification models, which are divided as clustering based methods [4], [13], [27] and domain adaptation models [5], [12], [16], [17], [26]. Clustering based person re-id models only analyze the unlabeled datasets and generally yield poor pedestrian matching performance due to the lack of strong supervised tuning and optimization. Nevertheless, Cross Domain Person Re-id (CDPR) models is a preferable solution to overcome the shortcomings in clustering methods. It learns transferable knowledge from a completely labeled dataset (denoted as source domain), and embed it into an unlabeled dataset (denoted as target domain), which is disparate within the objective variations and contains none overlapping

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

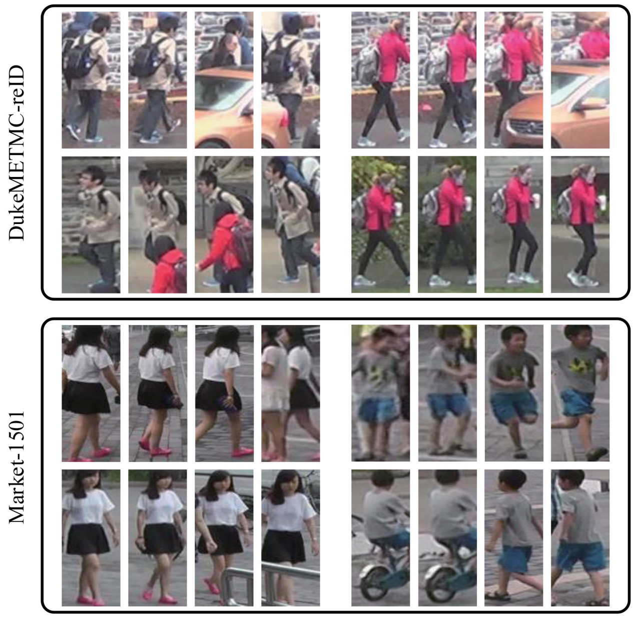

FIGURE 1. Several samples in DukeMTMC-reID and Market-1501 datasets. This figure illustrates the variations in background, illumination, viewpoints in different cameras and the domain gap in cross domain person re-identification task.

pedestrian identity with source domain. The CDPR model can resolve a series of urgent person re-identification tasks lacking of sufficient time to annotate vast identity labels.

In addition to relax the constraint of unlabeled data, CDPR model is up against more complex situations than conventional person re-id problem. The most challenging task in CDPR problem is to bridge the distribution gap between source and target domains (shown in Figure 1), which attracts much more research attention in existing CDPR approaches [16], [26]. Xu *et al.* [26] proposed an attribute feature learning based deep neural network to transfer augmented attribute feature across domains to resolve cross dataset person re-id task, so as to distinguish pedestrian with similar attributes by the additional learned extra image information. Lv *et al.* [16] employ a spatial-temporal information transfer mechanism to conduct unsupervised learning in target domain, which trained a robust classification network in source domain as the guidance to boost the discriminative ability of the person's feature extracted from target domain. After obtaining the final feature, it adopted learning-to-rank based boosting process to enhancively train the classification based model of unlabeled target domain.

Compared with clustering based person re-identification models, which often annotate soft labels on pedestrian images to provide identity information, CDPR models only focus on alleviating the domain gap between source and target domains through the trained matching model in source domain. They ignore the pedestrian identity information and performs weak effectiveness in realistic scene. Furthermore, directly introducing the soft-label method into target domain is not an effective way, because clustering based person re-id models have a conventional assumption that the number of pedestrian identities is acknowledged. This is an impractical assumption when we focus on target domain, and they can not estimate the cluster number to provide sufficient soft identity information.

To make up these drawbacks analyzed above, we propose a novel multi-task transfer learning framework for cross domain person re-identification, which is not only learning discriminative feature representation from source domain, but also transferring it into target domain with cluster estimating algorithm to support a soft multi-task learning procedure as well as in source domain, namely Deep Multi-task Transfer Network (DMTNet) method. It can achieve the unknown class number clustering soft-label process in unsupervised person re-identification. Our DMTNet is composed by an identity information classification module, attributes-attention module, and identification module in source domain, as well as identity information estimation module in target domain and a multi-task transfer module both in source and target domains, which guarantees the feature representative ability and identity estimation in the multi-task transfer learning course.

II. RELATED WORK

In this section, we review the research works on unsupervised person re-identification, which is divided as clustering based person re-id and cross domain person re-id methods according to surveys [9], [24]. In addition, this paper introduce a multi-task learning architecture into CDPR model, so we also discuss some multi-task learning approaches in supervised person re-identification.

A. CLUSTERING BASED PERSON RE-IDENTIFICATION

Existing clustering based person re-identification models often minimize the distance between similar pedestrian images and maximize the distance between dissimilar images by the soft labels provided by clustering algorithms [3], [13], [27]. Specifically, Lin *et al.* [13] focused on diversity across different identities and similarity within the same identity, and utilized a diversity regularization term in the bottom-up clustering procedure to balance data volume of each cluster, which achieved an effective trade-off between the diversity and similarity. Yang *et al.* [27] proposed a patch-based unsupervised learning framework in order to learn discriminative feature from patches instead of the whole images. The patch-based learning leveraged similarity between patches to learn discriminative model by an unsupervised patch-based discriminative feature learning loss and an image-level feature learning loss as the guidance. Ding *et al.* [3] introduced a statistic conception of 'dispersion' to constrain the clustering algorithm following the dispersion state and proposed a novel clustering based unsupervised person re-id models which can exploit the underlying feature space for unlabeled pedestrian image data.

Though these clustering based unsupervised person re-id models have achieved progressive improvements, they still keep a way from realistic application due to the lacking of any priori knowledge of the unlabeled data. The chief drawback of them is that they make the assumption that unlabeled person re-id dataset has a known identity number. That is less rigorous compared with application in real scene.

B. CROSS DOMAIN PERSON RE-IDENTIFICATION

Another frequently-used strategy is the cross domain person re-id solution. They are either leverage the feature gap or transform the image style to bridge the distribution gap [2], [10], [23]. Li *et al.* [10] proposed a pose disentanglement and adaptation network aiming at learning deep image representation with pose and domain information properly disentangled, and it can perform pose disentanglement across domains without supervision in identities. Chen *et al.* [2] proposed an instance-guided context rendering scheme for cross-domain person re-identification, which transfer the source person identities into diverse target domain contexts to enable supervised re-id model learning in the unlabeled target domain. Wei *et al.* [23] proposed a person transfer generative adversarial network on person transfer to bridge the domain gap among datasets, which consider extra constraints on the person foregrounds to ensure the stability of their identities during transfer.

These cross domain person re-id models can alleviate the domain gap by image style transfer or learning a shared feature space to perform competitive results, compared with clustering based method. However, they are highly relied on source domain when train a discriminative feature representation without consider identity information of target data.

C. MULTI TASK PERSON RE-IDENTIFICATION

Both clustering based and cross domain models have their weakness, thus we are prone to design a novel multi-task learning framework can not only estimate the cluster number in target domain to utilize identity information but also bridge the domain gap between source and target domains. It can sufficiently take advantage of the soft identity information during the cluster estimation to overcome the weakness in existing cross domain person re-identification models. In this subsection, we describe the existing multi-task learning based person re-identification models.

Existing multi-task person re-id models are almost under supervised framework [1], [14], [21]. Chen *et al.* [1] are the first to integrate a binary classification task and ranking task into a unified framework, named MTDnet. Inspired by MTDnet, Ling *et al.* [14] proposed a multi-task learning network with four different losses for person re-identification, including person re-identification, pedestrian identity task and pedestrian attribute task, who provide complementary information from different perspective. Wang *et al.* [21] proposed a multi-task attentional network with curriculum sampling method, which contains a fully attentional block and a curriculum sampling method for training ranking losses.

These multi-task learning approaches can integrate several task-specific goals into a unified network to boost the feature representation learning. Different with them, our DMTNet composes clustering, attribute learning and domain adaptation into a multi-task cross domain person re-identification approach, which can conduct unsupervised person re-identification without a prior-acknowledge pedestrian number in target domain. Detail description of our DMTNet is in Section III.

III. OUR APPROACH

In this section, we describe our proposed Deep Multi-task Transfer Network (DMTNet) in detail, and illustrate the optimization of the whole algorithm.

A. APPROACH OVERVIEW

Aiming at solving cross-domain person re-identification problem, we propose a Deep Multi-task Transfer Network (DMTNet), which is constructed by two main manifolds, including a identity information classification task, an attributes-attention task, identification task in source domain, and identity information estimation module, a multi-task transfer module in target domain. In novelty, we design an attributes-attention mechanism due to the positive affects produced by introducing attributes in existing person re-id methods. This module can learn the attribute importance for the learned attribute feature, and then synthesize all the attribute features combined with their importance to produce a final attributed feature. Then, we employ part of the source domain as the guidance of estimating pedestrian clustering numbers of target domain, which is in charge of a soft classification task for target data. Finally, the estimated target estimated soft labels and the attributed network are combined to conduct multi-task network training in the target domain with a transfer function. Therefore, our multi-task transfer module can bridge the domain gap between source and target domains, which is achieved by the attributes-attention and self-information estimation modules. Detail architecture can be seen in Figure 2.

B. MULTI-TASK LEARNING IN SOURCE DOMAIN

In cross domain person re-identification, here is the assumption that $D^s = \{(x_i^s, y_i^s)|, i = 1, \cdots, i, \cdots, N^s\}$ is the completely labeled source domain, where x_i^s is the i-th image in source domain with its correlated label y_i^s, and N^s is the image number in source domain. Furthermore, there is also a target unlabeled domain $D^t = \{x_i^t|, i = 1, \cdots, i, \cdots, N^t\}$, where contains the target matching data, x_i^t is the i-th image in target domain and N^t is the number of pedestrian images in target domain. Then, we introduce the attributes label in source domain to achieve the muti-task network, and define the attributes annotations $a_i = (a_i^1, \cdots, a_i^j, \cdots, a_i^m)$ for i-th image x_i^s in source domain, where a_i^j represents the j-th attribute for x_i^s. Note that, there is none attribute labels in target domain D^t, and we are aiming at learning attribute attention network in source domain and transferring it into target domain.

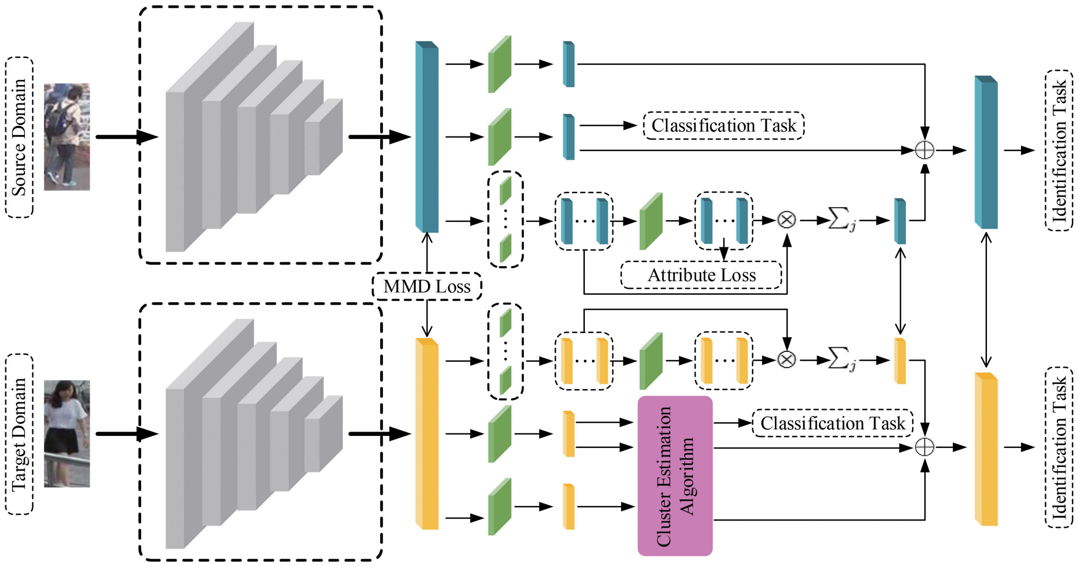

FIGURE 2. Architecture of our proposed deep multi-task transfer network. Both of source and target data is fed into a shared CNN module and it obtains their basic pedestrian feature $F_0^{s(i)}$, which is added a MMD constraint. Then, the identity information classification task, attributes-attention task and identification task is fixed on source domain, while identity information task and a multi-task transfer module are deployed in target domain. After all, this deep multi-task transfer network can output a fused feature for pedestrian representation by a concatenation layer. Besides, this architecture employs a novel identity estimating cluster algorithm to annotate soft labels, which is used to conduct identification training for target data.

Firstly, we design the multi-task network for source domain, to achieve identity-classification, attribute-attention, and identification tasks. These main processes are implemented by a backbone network to extract basic pedestrian features and a series of attention operations, both of which are constrained by related objective functions. In detail, a backbone feature extracting network $\gamma(x_i^s, \theta^0)$ is employed for source pedestrian image x_i^s, where $\gamma(\cdot, \theta^0)$ is the backbone convolutional neural network with parameters θ^0. Therefore, the basic pedestrian feature $F_0^{s(i)}$ for x_i^s can be obtained by the backbone network,

$$F_0^{s(i)} = \gamma(x_i^s, \theta^0) \tag{1}$$

After extracting basic feature, we add several parallel fully-connected layers to achieve feature transformation for pedestrian classification, attribute attention and identification features. They are expressed by $f(\cdot, \theta^{cl})$,$f(\cdot, \theta^{at})$ and $f(\cdot, \theta^{id})$ separately, where $\theta^{at} = \{\theta^{at(1)}, \cdots, \theta^{at(j)}, \cdots, \theta^{at(m)}\}$ is the collection of attribute layer parameters and $\theta^{at(j)}$ is the parameter of fully-connected layer for j-th attribute. Through these transformers, we can acquire the task-specific features for x_i^s in different tasks, including classification-specific feature $F_{cl}^{s(i)} = f(F_0^{s(i)}, \theta^{cl})$, attributes-specific features $F_{at}^{s(i)} = [F_{at(1)}^{s(i)}, \cdots, F_{at(j)}^{s(i)}, F_{at(m)}^{s(i)}]$, where $F_{at(j)}^{s(i)} = f(F_0^{s(i)}, \theta^{at(j)})$, and identification-specific feature $F_{id}^{s(i)} = f(F_0^{s(i)}, \theta^{id})$.

From the procedure of these three task-specific features, it is well acknowledge that the basic feature $F_0^{s(i)}$ contains much more information containing class, attributes, and identity, which are important for person re-identification. To strengthen the representative ability of the basic feature, we devise three objective loss functions for the task-specific features.

The first one is the classification loss for $F_{cl}^{s(i)}$, on where we employ the softmax layer to generate the probabilities belonging to each identity in source domain,

$$L_{cl} = -y_i^s \log(\text{softmax}(F_{cl}^{s(i)})) \tag{2}$$

This item can classify each image into pedestrian classes, and make the basic feature preserve pedestrian identity characteristic information.

Moreover, our DMTNet method introduces an attribute attention loss function on attribute-specific features to retain the attribute information in pedestrian basic features. For the attributed-specific feature $F_{at}^{s(i)}$, we assume an attention matrix $U_{at}^{s(i)} = f(F_{at}^{s(i)}, \theta^u)$, which estimate the importance of each attribute. This part is achieved by the constraint between the attribute labels and the attention matrix,

$$L_{at} = \sum_{i=1}^{N^s} \sum_{j=1}^{m} \| U_{at}^{s(i)} - U_a^{s(i)} \| \tag{3}$$

where $U_a^{s(i)} \in \mathbb{R}^{m \times m}$ is a composition of the pedestrian attribute labels, which is expanded into a column in $1 \times m$. Each element in the column is in [0, 1] and denotes whether the pedestrian contain the attribute. From this constraint, the attribute attention coefficient of each pedestrian image can be obtained to strengthen the representative ability for the attribute-specific feature,

$$\overline{F_{at}^{s(i)}} = \sum_J F_{at}^{s(i)} \times U_{at}^{s(i)} \tag{4}$$

where $\sum_J$ denotes the column sum calculation.

Finally, the person re-identification is a matching problem in essence. We have acquire the identification-specific feature for each pedestrian image. However, the matching feature not only require the identification-specific feature, but also contain the pedestrian identity information and its attribute information. Therefore, we conclude these three task-specific features into a fused feature representation,

$$F_{mat}^{s(i)} = [F_{id}^{s(i)}, F_{cl}^{s(i)}, \overline{F_{at}^{s(i)}}] \tag{5}$$

where $[\cdot]$ is the concatenation.

After the final feature representation for each pedestrian image, we employ the triplet loss function in source domain. For a positive pedestrian image pair, we generate a triplet samples, including a positive pair (x_i^s, x_j^s) and a negative sample x_k^s). In this triplet function, we select the negative sample by its similarity to positive sample pair in first rank. Their final feature representations should following the expected state as FaceNet [20],

$$\left\| F_{mat}^{s(i)} - F_{mat}^{s(j)} \right\|_2^2 + \alpha < \left\| F_{mat}^{s(i)} - F_{mat}^{s(k)} \right\|_2^2 \tag{6}$$

Therefore, we introduce the triplet loss function to ensure the matching performance,

$$L_{mat} = \sum_{i}^{N} \left[\left\| F_{mat}^{s(i)} - F_{mat}^{s(j)} \right\|_2^2 - \left\| F_{mat}^{s(i)} - F_{mat}^{s(k)} \right\|_2^2 + m \right]_+ \tag{7}$$

For the devised feature extracting mechanism and multi-task learning loss functions, our DMTNet can conduct a matching procedure in source domain. After that, the main left problem is how to transfer this network into target domain, and how to implement the multi-task learning in the completely unlabeled target domain.

C. MULTI-TASK LEARNING IN TARGET DOMAIN

In the last subsection, we principally describe the DMTNet model in source domain, and this part is prone to build the multi-task transfer model to clinically solve the multi-task learning in target domain.

In target domain, three specific tasks are taken into account to implement. DMTNet focuses on transferring the multi-task learning from source domain, including classification, attribute learning and identification tasks. The most intractable transfer task is the classification, because the source and target domains do not share any common pedestrian. The attribute learning and identification tasks can be alleviate the domain gap by Maximum Mean Discrepancy (MMD) constraint [7].

Based on the pre-trained model from source domain, DMTNet can output a basic feature $F_0^{t(i)}$, classification-specific feature $F_{cl}^{t(i)}$, attribute-specific feature $F_{at}^{t(i)}$, and identification-specific feature $F_{id}^{t(i)}$ when input a pedestrian image x_i^t in target domain. For these task-specific features, we can bound the MMD constraint on the basic feature, attribute-specific feature and identification-specific feature to make distributions of source and target are consistency,

$$\begin{aligned} L_{mmd} &= \| \frac{1}{N^s} \sum_{i=1}^{N^s} F_0^{s(i)} - \frac{1}{N^t} \sum_{i=1}^{N^t} F_0^{t(i)} \|^2 \\ &\quad + \| \frac{1}{N^s} \sum_{i=1}^{N^s} \overline{F_{at}^{s(i)}} - \frac{1}{N^t} \sum_{i=1}^{N^t} \overline{F_{at}^{t(i)}} \|^2 \\ &\quad + \| \frac{1}{N^s} \sum_{i=1}^{N^s} F_{mat}^{s(i)} - \frac{1}{N^t} \sum_{i=1}^{N^t} F_{mat}^{t(i)} \|^2 \end{aligned} \tag{8}$$

Note that the MMD loss function do not constrain classification-specific feature thanks to it may lose some important identity information. It is better to train the classification task independently, but conservative unsupervised person re-identification clustering algorithm is based on exclusive pedestrian number, which can not be acknowledged previously in realistic scene. Thus, DMTNet proposes a Novel Identity Estimating Cluster (NIEC) algorithm to discover pedestrian identity clustering number and annotates soft labels on target data, which are used to train the classification task.

In Novel Identity Estimating Cluster algorithm, we are aiming at discovering new pedestrian identity by the experience gained from source domain. NIEC algorithm is inspired by Deep Embedding Clustering (DEC) [25]. However, the goal of NIEC method is not only determine the cluster points, but also to discover the number of clusters in target domain.

Following DEC approach, let $P_s(c^s|i)$ be the probability of source pedestrian image x_i^s belonging to identity cluster $c^s \in \{1, \cdots, C^s\}$, where C^s denotes the identity number of source pedestrian images. DEC employ a Student's t distribution as the initial parameterization,

$$P_s(c^s|i) \propto \left(1 + \frac{\left\| f_{cl}^{s(i)} - \mu_c \right\|^2}{\alpha} \right)^{-\frac{\alpha+1}{2}} \tag{9}$$

where $\mu_{c^s} \in \{\mu_{c^s}, c^s = 1, \cdots, C^s\}$ is the c-th identity cluster.

Assuming that pedestrian data indices are sampled uniformly (i.e. $P_s(i) = 1/C^s$), the joint distribution can be written as $P_s(i, c^s) = p(c^s|i)/C^s$. In target domain, we suppose the pedestrian data also following the t distribution, denoted as $P_t(c^t|i)$, where the $c^t \in \{1, \cdots, C^t\}$, where C^t denotes the identity number of target pedestrian images.

Because the classification task outputs the probability of each pedestrian image, the optimal solution to neutralize the domain distribution gap of the predicted probability is minimizing the KL divergence between joint distributions $P_s(i, c^s) = P_s(c^s|i)/C^s$ and $P_t(i, c^t) = P_t(c^t|i)/C^t$. We employ the symmetrized version of KL-divergence,

$$\begin{aligned} L_{KL} &= KL(P_s||P_t) + KL(P_t||P_s) \\ &= \sum_{i=1}^{N^s} P_s(i) \ln \frac{P_s(i)}{P_t(i)} + \sum_{i=1}^{N^t} P_t(i) \ln \frac{P_t(i)}{P_s(i)} \end{aligned} \tag{10}$$

This item can keep the distribution gap between source and target domains in consistency, but it needs to know the explicit identity number of target domain. In realistic cross domain person re-identification, it is very hard to know the pedestrian number for target data. Therefore, DMTNet devises an identity number estimation mechanism to seek for person account in target domain, which can transfer the classification model learned from source domain.

Through the MMD and KL loss functions, the pedestrian images both in source and target domains are transformed into a shared feature space. That makes the identity number estimating model trained by source data is appropriate for target data. According to this theoretical basis, we utilize the source data to train our NIEC algorithm. In detail, we split the C^s known classes in D^s into a training subset D^s_t with C^s_t classes, and a validating subset D^s_v with $C^s_v = C^s - C^s_t$ pedestrian classes($C^s_t : C^s_v = 4 : 1$). Moreover, the target data in D^t is regarded as testing subset.

Then we run a semi-supervised k-means clustering method on $D^s_t \cup D^t$ to estimate the number of identity in D^t. Namely, during k-means, we force images in the training subset D^s_t to map to clusters following their ground-truth labels, while images in the validation subset D^s_v are considered as additional "unlabeled" data. We launch this constrained k-means multiple times by sweeping the number of total categories C in $D^s_t \cup D^t$, and measure the constrained clustering quality on $D^s_t \cup D^t$ by the estimation in D^s_v. To this end, we employ two evaluating criteria [6] on the estimating result of D^s_v to evaluate the clustering effectiveness under classes C.

The first criterion is Overall Clustering Accuracy (ACC), which is applicable to the C^s_v labeled classes in the validation subset D^s_v, and it is given by,

$$\max_{g\in \mathrm{Sym}(N^s_v)} \frac{1}{N^s_v} \sum_{i=1}^{N^s_v} 1\left\{\bar{y}^s_i = g\left(y^s_i\right)\right\} \tag{11}$$

where N^s_v is the number of validating images, and $g(y^s_i)$ denotes the ground-truth label, while the $\bar{y}^s_i$ is the estimated clustering assignment for each image in D^s_v. This term ensures to estimate correct labels as much as possible for validating subset D^s_v.

The another criterion is Cluster Metric Measurement (CMM) by capturing notions of intra-identity cohesion and inter-identity separation based on estimated clusters, which is applicable to the unlabeled data D^t. This constraint is according to,

$$\sum_{f^{t(i)}_{cl}\in D^t} \frac{b(f^{t(i)}_{cl}) - a(f^{t(i)}_{cl})}{\max\{a(f^{t(i)}_{cl}), b(f^{t(i)}_{cl})\}} \tag{12}$$

where $f^{t(i)}_{cl}$ is the classification-specific feature point in target domain, $a(f^{t(i)}_{cl})$ is the average distance between $f^{t(i)}_{cl}$ and all other data samples within the same cluster, and $b(f^{t(i)}_{cl})$ is the smallest average distance of $f^{t(i)}_{cl}$ to all points in any other cluster (of which $f^{t(i)}_{cl}$ is not a member).

```
**Algorithm 1** Deep Multi-Task Tranfer Network (DMTNet)
**Initialization:** The backbone network $\gamma(x^s_i, \theta^0)$, and the task-specific feature extractors $f(\cot, \theta^{at})$, $f(\cot, \theta^{cla})$ and $f(\cot, \theta^{id})$ by the data in labeled source domain $D^s = \{(x^s_i, y^s_i)|, i = 1, \cdots, i, \cdots, N^s\}$. Given an evaluated identity number $C^0 \leq C^* \leq C^{max}$ for $D^s_t \cup D^t$, and parameter $m = 0.35$ in Eq. 7.
**Training in source domain:**
**for** $t \in 1, \cdots, N_s$ **do**
    Train $\theta^0, \theta^{at}, \theta^{cla}, \theta^{id}$ on $D^s$ by Eq.2, 3 and 7.
**end for**
**for** $C^0 \leq C^* \leq C^{max} do$
    K-means cluster and annotate soft labels for the target data.
    **Training in target domain:**
    **for** $t \in 1, \cdots, N_t$ **do**
        Train $\theta^0, \theta^{at}, \theta^{cla}, \theta^{id}$ on $D^s$ by Eq.2, 3, 7, 8, and 10 for the target data.
        Train $\theta^0, \theta^{at}, \theta^{cla}, \theta^{id}$ on $D^s$ by Eq.2, 3 and 7 for the source data.
        Train the whole network by Eq.11, and 12.
        **If** the error is convergence, **Return** $\theta^0, \theta^{at}, \theta^{cla}, \theta^{id}$.
    **end for**
    Select the optimal value of $C*$.
**end for**
**Training in target domain again by the** $C^*$.
**Return** the parameters of DMTNet.
```

When these two criteria are in convergence under a fixed value, the cluster number C^* in $D^s_t \cup D^t$ can be obtained, and the identity number is $C^* - C^s_t$. Through the obtained clusters in target domain, we can annotate soft identity label for each pedestrian image in target domain D^s. Thus, the whole multi-task learning approach can be transferred into target data to implement and achieve the fused feature $F^{t(i)}_{mat}$. The testing procedure between probe and gallery in target domain can be conducted by,

$$c^t_i = \arg\min_c \|F^{t(i)}_{mat} - F^{t(c)}_{mat}\|^2, \quad 1 \leq c \leq C^t \tag{13}$$

where c^t_i is the predicting label given a pedestrian image in target domain. Our DMTNet algorithm is also concluded in Algorithm 1.

IV. EXPERIMENTS

A. DATASETS AND EXPERIMENTAL SETTINGS

To validate the efficacy of the DMTNet approach, we implement several evaluating experiments on two large scale datasets, DukeMTMC-reID [19], [29] and Market-1501 [28], which have attribute labels and are widely used in cross domain person re-identification models.

DukeMTMC-reID dataset [19], [29] is a sufficient large scale person re-identification dataset, which is suitable for deep neural network training task. The pedestrian images are captured by 8 surveillance cameras, and they are composed by 36,411 annotated images with 1,404 persons.

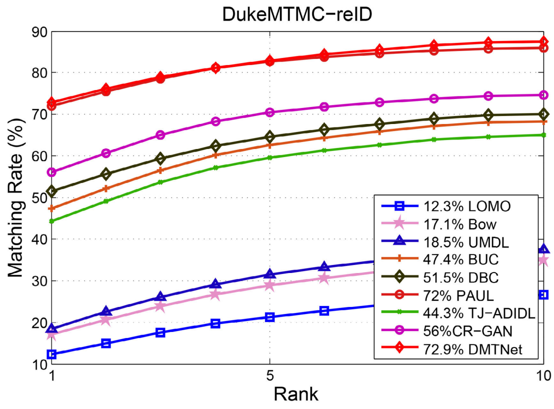

FIGURE 3. CMC performance on DukeMTMC-reID dataset.

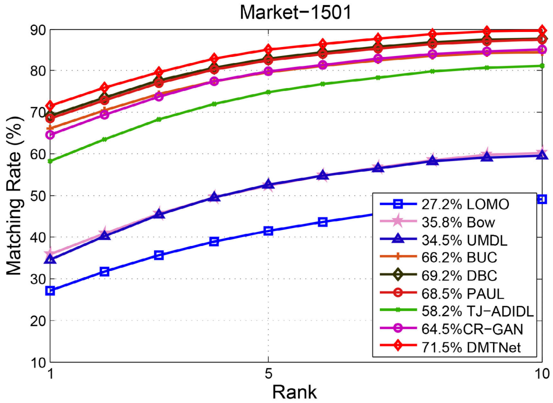

FIGURE 4. CMC performance on Market-1501 dataset.

They are formed by 702 person with 16,522 training images, and 2,228 probe pedestrian images belonging to other 702 persons, which has 17,661 gallery images.

Market-1501 dataset [28] contains 32,668 annotated pedestrian images belonging to 1,501 persons, which are captured by 6 surveillance cameras. This large scale dataset is divided as two partitions by their purpose, where the one is utilized for training with 12,936 images from 751 individuals and the left 19,732 images from 750 individuals are adopted for testing. There are 3368 probe images from the 750 testing persons are employed to matching target identities in the gallery.

Evaluation protocols are kept in consistency with conventional person re-id models. We utilize the Cumulative Matching Curve (CMC) to produce the ranking accuracy, and adopt the mean Average Precision (mAP) to evaluate the performance of our approach, which reflects the overall precision and recall rates. In this section, we describe the rank-1,rank-5, and rank-10 and mAP to show the performance on these two experimental datasets (Table 3), and draw the CMC curves as the performance of rank-n ($1 \leq n \leq 20$)(Figure 3, 4).

B. IMPLEMENTATION DETAIL

To train our deep multi-task transfer network, we introduce the ResNet50 with pre-trained parameters on ImageNet as the basic feature extractor following Luo *et al.* [15] and use Pytorch to achieve the network. The objective function of DMTNet is optimized by Adam solver [8] with a mini-batch of 32 on Ubuntu 16.04 system with NVIDIA GeForece GTX 2080Ti GPU. The learning rate of the whole network is initialized by $2e$-4 when is training a half process, and then will be decayed to 0 at the end of training. The parameter m is set to be 0.35, and the dimentsion of output matching feature is 256.

C. PERFORMANCE EVALUATION

1) COMPARED METHODS

To show the effectiveness on these two datasets, we choose three hand-craft feature based machine learning approaches (LOMO [11], BoW [28], and UMDL [18]), several deep clustering based methods (BUC [13], DBC [3] and PAUL [23]), and two state of the art cross domain approaches (TJ-AIDL [22], and CR-GAN [2]).

Specifically, LOMO [11] is an effective hand-craft feature representation of local maximal occurrence, which analyze the horizontal occurrence of local features and maximizes the occurrence to make a stable representation against viewpoint changes. This feature is often utilized into conventional machine learning models to solve person re-identification. BoW [28] is a Bag-of-Words model, which accommodates local features and enables fast global feature matching. UMDL [18]) is a asymmetric multi-task dictionary learning model to learn view-invariant and identity-discriminative information from unlabeled target data. For deep clustering based methods, BUC [13] is a bottom-up clustering approach to jointly optimize a convolutional neural network and the relationship among the individual samples. DBC [3] is a novel clustering based unsupervised person re-id models which can exploit the underlying feature space for unlabeled pedestrian image data by the statistic concept of 'dispersion'. PAUL [27] is a patch-based unsupervised learning framework in order to learn discriminative feature from patches instead of the whole images, combined with an unsupervised patch-based discriminative feature learning loss. Furthermore, the two state-of-the art cross domain approaches (TJ-AIDL [22], and CR-GAN [2]) also conducted experiments on Market-1501 and DukeMTMC-reID. TJ-AIDL [22] transfers the labeled information of an existing dataset to a new seen unlabeled target domain for person re-id without any supervision in the target domain, which simultaneously learns an attribute-semantic and identity discriminative feature representation space transferrable to the target domain. CR-GAN [2] formulates a dual conditional generative adversarial network that augments each source person image with rich contextual variations, and leverages abundant unlabeled target instances as contextual guidance for image generation.

2) PERFORMANCE ON DukeMTMC-reID

For DukeMTMC-reID dataset, we utilize the Market-1501 as the source, and DukeMTMC-reID is the target domain, which

TABLE 1. Performance (%) comparison on DukeMTMC-reID dataset.

Dataset	Source: Market-1501, Target: DukeMTMC-reID			
Models	rank-1	rank-5	rank-10	mAP
LOMO [11]	12.3	21.3	26.6	4.8
Bow [28]	17.1	28.8	34.9	8.3
UMDL [18]	18.5	31.4	37.6	7.3
BUC [13]	47.4	62.6	68.4	27.5
DBC [3]	51.5	64.6	70.1	30.0
PAUL [23]	72.0	82.7	86.0	53.2
TJ-AIDL [22]	44.3	59.6	65.0	23.0
CR-GAN [2]	56.0	70.5	74.6	33.3
DMTNet(Ours)	72.9	83.0	87.5	53.8

is set as the compared cross domain methods' setting. Table 1 reports the comparison between our DMTNet approach and these compared methods. From this table, it can be seen that our DMTNet model obtains rank-1 accuracy of 72.9%, rank-5 accuracy of 83.0%, rank-10 accuracy of 87.5% and mAP rate of 53.8%, and its CMC curve in drawn in Figure 3. These results are superior to most of recent approaches, and have a important significance for cross domain person re-identification.

Compared with hand-craft feature based machine learning methods, our DMTNet can extract more discriminative information according to different tasks, and leaves them a large margin in rank-*n* accuracies and mAP rate. Contrast to clustering based unsupervised person re-identification models, our approach can estimate the cluster points for conducting multi-task learning in target domain (improve at least 0.9% (72.9-72.0) of rank-1 accuracy). Compared to cross domain models, our method preserves the identity information in multi-task learning procedure, and it improves 16.9% (72.9-56.0) in rank-1 accuracy and 20.5% (53.8-33.3) in mAP. Moreover, their CMC curves of the compared models are drawn in Figure 3.

3) PERFORMANCE ON MARKET-1501.

For this dataset, we choose the DukeMTMC-reID as the source domain, and the cross domain person re-id models keep in consistency with this setting. We report the results in Table 2, and it shows the superiority of our proposed, compared with hand-craft feature based machine learning, clustering and domain adaptation based person re-id models. DMTNet conduct experiments and obtain rank-1 accuracy of 71.5 and mAP rate of 42.3, which surpass the baselines and state of the art methods, their CMC curves also shows this result (Figure 4). From the performance on two datasets, our proposed cross domain person re-identification model using multi-task transfer learning framework is proved to outperform existing unsupervised person re-identification methods.

D. ABLATION AND DISCUSSION

This paper proposes a novel multi-task transfer network, integrating classification, attribute learning, and identification tasks into a unified framework, and design a cluster estimating algorithm for target domain. In this part, we will evaluate the influence of them, and make several discussions of our DMTNet.

TABLE 2. Performance (%) comparison on Market-1501 dataset.

Dataset	Source: DukeMTMC-reID, Target: Market-1501			
Models	rank-1	rank-5	rank-10	mAP
LOMO [11]	27.2	41.6	49.1	8.0
Bow [28]	35.8	52.4	60.3	14.8
UMDL [18]	34.5	52.6	59.6	12.4
BUC [13]	66.2	79.6	84.5	38.3
DBC [3]	69.2	83.0	87.8	41.3
PAUL [23]	68.5	82.4	87.4	40.1
TJ-AIDL [22]	58.2	74.8	81.1	26.5
CR-GAN [2]	64.5	79.8	85.0	33.2
DMTNet(Ours)	71.5	85.0	89.7	42.3

TABLE 3. Performance comparison in different tasks.

Dataset	DukeMTMC-reID		Market-1501	
Models	rank-1	mAP	rank-1	mAP
MTwAANet	65.5	44.6	60.6	32.7
MTwCNet	68.6	48.1	66.2	38.3
MTwINet	45.6	27.0	56.2	33.5
MTwNNet	73.3	54.0	72.5	42.8
DMTNet	72.9	53.8	71.5	42.3
DMTNet+ReRanking	74.1	55.7	76.0	47.9

1) EVALUATION OF ATTRIBUTE ATTENTION TASK

We employ attribute learning as one of the multi tasks due to the effectiveness of existing attribute based person re-id models, and develop it by our attribute attention subnet to supplement the attribute-specific information in final matching features. As a comparison, we remove this attention task and combine the attribute feature directly with other tasks both in source and target domains, named as Multi Task Without Attribute Attention (MTwAANet).

Table 3 shows the results of MTwAANet, and it achieves rank-1 accuracy of 65.5% on DukeMTMC-reID and 60.6% on Market-1501, which is lower than the performance than our DMTNet as well as mAP performance. This comparison demonstrates the positive effect of the attribute attention task not only on rank-1 performance but also on mAP criterion.

2) EVALUATION OF CLASSIFICATION TASK

In existing cross domain person re-identification methods, the classification task is often employed in source domain rather than target because target domain is lack of annotations which can not support the classification. We attempt to estimate cluster number and annotate a soft class labels to each image in target domain. Thus, we make it possible to integrate classification task into target domain, and preserve soft identity information on the matching feature for each image in target domain. To evaluate this task, we eliminate this term in the whole network, named as Multi Task without Classification (MTwCNet).

We can find that the MTwCNet obtains a rank-1 accuracy of 68.6%(66.2%) on DukeMTMC-reID(Market-1501) dataset, and mAP of 48.1%(38.3%). The superiority of our DMTNet is at least 4.3% of rank-1 accuracy, and the importance of the classification shows the identity information is a significant component for pedestrian feature matching.

3) EVALUATION OF IDENTIFICATION TASK

The identification task is in charge of the basic matching ability, which assimilate the identity and attribute information to constitute the matching feature. We remove this basic feature, and direct fuse the classification-specific feature and attribute-specific feature to conduct final matching process, named Multi Task without Identification (MTwINet), to evaluate this basic feature extraction.

The identification task reveals its decisive effect on the pedestrian feature matching procedure through the performance comparison between MTwINet and our DMTNet. The difference between them is at least 15.3%, which is the largest discrepancy in every task.

4) DISCUSSION OF CLUSTER NUMBER ESTIMATION

Our DMTNet approach proposes a cluster number estimating algorithm, which use KL divergence to leverage the distribution gap between source and target domains for classification task, and learn cluster knowledge from source domain to estimate target cluster number. Because it is always unknown of the identity number of unsupervised person re-identification in realistic, this strategy can resolve this problem. To validate the cluster number estimating algorithm, we set a constant identity number from the real account of the target domain as a comparison, instead of estimated number, which is named as Multi Task with identity Number (MTwNNet).

With the guidance of real identity number in target domain, MTwNNet achieves the rank-1 accuracy of 73.3%(72.5%) on DukeMTMC-reID(Market-1501), which is higher than our original DMTNet. The distance between them is only [0.4%,1%], and it shows that our DMTNet can resolve the problem of lacking target identity number in realistic, while retains little distance between MTwNNet.

5) EVALUATION OF METRICS

To evaluate the evaluating metric of the pedestrian feature matching, we adopt re-ranking [30] technology to improve the performance of our DMTNet. Following the parameter setting of re-ranking, our method improves a margin of 1.2%(4.5%) on DukeMTMC-reID(Market-1501), which is the DMTNet+ReRanking in Table 3. That illustrates our DMTNet has a prospect of improvement when combine different metrics or complex feature extractors.

V. CONCLUSION

In this paper, we present a novel multi-task transfer network for cross domain person re-identification. This approach aims to solve the target identity information preserving and target cluster number estimating problem, by a soft classification task and identity cluster estimating algorithm. It can not only learn discriminative feature representation from source domain, but also transfer them into target domain with cluster estimation to support a soft multi-task learning procedure as well as source domain. Furthermore, extensive experiments demonstrate the effectiveness of our proposed DMTNet method.

REFERENCES

[1] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–17.

[2] Y. Chen, X. Zhu, and S. Gong, "Instance-guided context rendering for cross-domain person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 232–242.

[3] G. Ding, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Towards better validity: Dispersion based clustering for unsupervised person re-identification," 2019, *arXiv:1906.01308*. [Online]. Available: https://arxiv.org/abs/1906.01308

[4] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Computing, Commun., Appl. (TOMM)*, vol. 14, no. 4, p. 83, 2018.

[5] A. Genş and H. K. Ekenel, "Cross-dataset person re-identification using deep convolutional neural networks: Effects of context and domain adaptation," *Multimed Tools Appl*, vol. 78, no. 5, pp. 5843–5861, Mar. 2019.

[6] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8401–8409.

[7] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5576–5588, Dec. 2016.

[8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[9] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[10] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C. F. Wang, "Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7919–7929.

[11] 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. (CVPR)*, Jun. 2015, pp. 2197–2206.

[12] S. Lin, H. Li, C.-T. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," 2018, *arXiv:1807.01440*. [Online]. Available: https://arxiv.org/abs/1807.01440

[13] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom–up clustering approach to unsupervised person re–identification," in *Proc. AAAI*, vol. 33, Aug. 2019, pp. 8738–8745.

[14] H. Ling, Z. Wang, P. Li, Y. Shi, J. Chen, and F. Zou, "Improving person re-identification by multi-task learning," *Neurocomputing*, vol. 347, pp. 109–118, Jun. 2019.

[15] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person Re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 1–9.

[16] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross–dataset person re-identification by transfer learning of spatial–temporal patterns," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7948–7956.

[17] J. Lv and X. Wang, "Cross-dataset person re-identification using similarity preserved generative adversarial networks," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.* Cham, Switzerland: Springer, 2018, pp. 171–183.

[18] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised Cross–Dataset Transfer Learning for Person Re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1306–1315.

[19] 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.* Cham, Switzerland: Springer, 2016, pp. 17–35.

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

[21] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 365–381.

[22] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute–identity deep learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.

[23] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[24] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, and D.-S. Huang, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, pp. 354–371, Apr. 2019.

[25] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[26] B. Xu, J. Liu, X. Hou, K. Sun, and G. Qiu, "Cross domain person re–identification with large scale attribute annotated Datasets," *IEEE Access*, vol. 7, pp. 21623–21634, 2019.

[27] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3633–3642.

[28] 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. (ICCV)*, Dec. 2015, pp. 1116–1124.

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

[30] 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. (CVPR)*, Jul. 2017, pp. 1318–1327.

HUAN WANG received the master's degree and the Ph.D. degree in system analysis and integration from Yunnan University, in 2008 and 2013, respectively. She is currently an Associated Professor with the Baoji University of Arts and Sciences. Her research interests include deep learning, complex networks, complex systems, and pattern recognition research.

JINGBO HU received the master's degree in computer architecture from Yunnan University, in 2010. He is currently an Associated Professor with the Baoji University of Arts and Sciences. His research interests include target detection, object recognition, and complex system research.

• • •