NPLP: A Noisy Pseudo-Label Processing Approach for Unsupervised Domain-Adaptive Person Re-ID

1st Tianbao Liang[1], 2nd Jianming Lv[12]*, 3rd Hualiang Li[3], 4th Yuzhong Liu[3]
[1]*School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China*
[2]*Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education, Guangzhou, Guangdong, China*
[3]*Key Laboratory of Occupational Health and Safety of Guangdong Power Grid Co., Ltd., Electric Power Research Institute of Guangdong Power Grid Co., Ltd., Guangzhou, Guangdong, China*
tbliang.gm@gmail.com, jmlv@scut.edu.cn, Li-hualiang@163.com, 411730335@qq.com

***Abstract*—Most of the existing unsupervised cross-domain person re-identification (re-ID) methods utilize pseudo-labels estimation to cast the unsupervised problem into a supervised problem, whose performance is limited by the quality of pseudo-labels. To address the problem, we propose a noisy pseudo-label processing (NPLP) approach to suppress the pseudo-labels noise and improve the performance of the person re-ID model. Specifically, we first summarize two types of pseudo-label noise that lead to the collapse of the re-ID model, as defined as mixed noise and fragmented noise. Secondly, we propose a different method which is composed of Startup Stage and Correcting Stage for pseudo-labels estimation to relieve these two types of noise respectively. The Startup Stage aims to decrease the ratio of the fragmented noise by increasing the recall of the clustering results. At the Correcting Stage, we evaluate the quality of the pseudo-labels and correct those low-quality pseudo-labels to suppress the mixed noise and generate more reliable pseudo-labels for the re-ID model to learn. At last, we build a feature learning strategy for unsupervised re-ID task and learn from the de-noised pseudo-labels iteratively. Extensive evaluations on three large-scale benchmarks show that the NPLP is competitive with most state-of-the-art unsupervised re-ID methods.**

I. Introduction

Person re-identification aims to retrieve all images of the target person from the surveillance videos collected from multiple different cameras. Recently, most person re-ID algorithms [1]–[3] have achieved impressive performance with large-scale labeled data. However, obtaining labeled data for person re-ID is time-consuming and expensive, so that such supervised methods have limited scalability and usability in real-world applications. How to effectively learn a model that can extract discriminative features for person re-ID on massive unlabeled data has been a challenging problem.

Some methods [4]–[6] have been proposed to tackle the unsupervised person re-ID problem without any labeled data. However, their performance is typically poor due to the lack of supervised signal, and thus being less effective for practical usage. More recently, many unsupervised cross-domain methods [7]–[16] have been proposed to utilize not only the unlabeled data (sampled from target domain) but also the outer labeled data (sampled from source domain) which could guide the network to learn from a more plausible distribution.

*Corresponding author

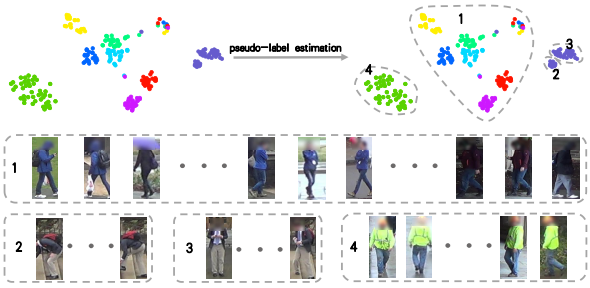

Fig. 1. T-SNE visualization of the initial feature embeddings on a part of DukeMTMC-reID. Points of the same color represent images of the same identity while points in the same circle means images that share the same cluster id. There are mainly two types of noise in pseudo-label estimation based methods: mixed noise (the cluster "1") and fragmented noise (the cluster "2" and "3"). The cluster "4" means the samples that can be easily grouped due to their huge difference with others.

The majority of the above unsupervised cross-domain methods [7], [9], [13], [15], [16] pre-train the re-ID model with the labeled source domain data and then adapt the model to the target domain by learning with pseudo-labels iteratively. They extract features of unlabeled data using the pre-trained model and assigned each sample a pseudo-label using unsupervised clustering methods (e.g., k-means and DBSCAN [17]). However, their performance is highly dependent on the quality of pseudo-label, and the mistake pseudo-labels may lead to the collapse of the re-ID model.

As shown in Fig 1, there are mainly two types of mistakes produced by the pseudo-label estimation procedure, as defined as mixed noise and fragmented noise respectively. The mixed noise means assigning the same pseudo-label to different persons' images, while the fragmented noise means splitting a person's images into different clusters. One reason for the production of these two types of noises is the intra-domain image style variations caused by different camera configurations [10]. Images with a similar appearance in the same camera may be assigned the same pseudo-label causes mixed noise. Meanwhile, the same person's images in different cameras may be assigned different pseudo-labels, leading to the generation of fragmented noise.

To relieve these two types of pseudo-label noises, we

propose a noisy pseudo-label processing (NPLP) approach to improve the performance of unsupervised cross-domain person re-ID task by generating higher quality pseudo-labels. Specifically, NPLP is a two-stage method and each stage suppresses one kind of noise. At the first stage, we suppress fragmented noise by increasing the probability that the same person's images captured by different cameras can be grouped. After doing so, basically all images of the same person can be divided into the same cluster, so the fragmented noise can be suppressed. Though the mixed noise still exists, images with different pseudo-labels have identifiable characteristics, which could weakly guide the re-ID model to learn useful knowledge of the target domain. As shown in Fig.1, the appearance of images in the cluster "1" are quite different from those in the cluster "2", "3" and "4". Therefore we call the first stage as *Startup Stage*. To further relieve the mixed noise, we propose the *Correcting Stage* based on a self-assessment (SA) score to evaluate the quality of each cluster and find out the noisy clusters. We then further perform pseudo-label estimation within those noisy clusters with low SA scores to correct the pseudo-labels. The experiments show that the performance of the re-ID model boosts significantly with these two stages running iteratively.

Furthermore, for the unsupervised feature learning, we introduce a hyper-network (HN) to encode multi-granularities features to take advantages of discriminative information from local to global and improve the testing efficiency.

Our contributions are as follows:

- We have summarized two types of pseudo-label noises in pseudo-label estimation based unsupervised domain adaptive person re-ID methods, namely, mixed noise and fragmented noise.
- A noisy pseudo-label processing approach (NPLP) is proposed to generate high quality pseudo-labels for re-ID model to learn.
- Extensive experiments are conducted on three large-scale benchmarks, and the results demonstrate that our simple and effective method is competitive with most start-of-the-art unsupervised domain adaptive person re-ID methods.

II. Related Work

a) Pseudo-Label Based Unsupervised Person re-ID: Recently, many methods are proposed to tackle unsupervised person re-ID in a self-training manner [4], [7], [9], [15], [16]. They repeated two steps, assigning pseudo-labels to training samples and training re-ID model with pseudo-labels until convergence. PUL [7] performs clustering on the unlabeled dataset and selects samples that are close to the cluster centroids to train the re-ID model. SSG [9] splits feature map into different parts and estimates pseudo-labels for each part independently and then trains the re-ID model iteratively. Meanwhile, BUC [4] proposes a bottom-up clustering method to group similar samples into the same identity iteratively. PAST [16] consists of conservative stage and promoting stage, while the former aims to improve feature representations, and the latter aims to explore the global distribution of unlabeled data. However, the performance of these methods is highly dependent on the quality of pseudo-labels, and noise in pseudo-labels may lead to the collapse of the re-ID model. Unfortunately, few methods consider the problem of pseudo-label noise as far as we know.

b) Deep Learning with Noisy Label: The problem of label noise has been widely studied in deep learning. Some methods [18], [19] estimate a noise transition matrix to correct the noisy labels. Some methods [20]–[22] design noise-robust loss to learn a noise-robust model. Literatures [23]–[25] try to filter out noisy labels by training iteratively and then learn from the clean labels. To the best of our knowledge, few methods have been proposed to address the problem of label noise in re-ID. DistributionNet [26] models feature uncertainty by adding extra embedding network and designs a loss to allocate uncertainty across training samples unevenly. BUC [4] makes assumptions about the number of images of each identity and designs a diversity regularization to avoid pseudo-label noise to some extent. MMT [27] utilizes auxiliary models to softly refined the pseudo-labels. Our method analyzes two types of pseudo-label noise in pseudo-label based methods, and designs a simple but effective training strategy to generate high quality pseudo-labels and then learn with them.

c) Part Based Person re-ID: Recent works [1]–[3] have shown the effectiveness of using multi-granularity features to construct robust features for supervised person re-ID. Similarly, [9], [14], [16], [28] utilized multi-scale features to tackle unsupervised person re-ID. [14] leverages similarity between patches to exploit the part affinity between instances. [9] estimates pseudo-label for each part independently and trained the re-ID model with the obtained pseudo-labels. EANet [28] proposes Part Aligned Pooling and Part Segmentation to enhance feature alignment. [16] adopts a similar strategy to [1] to construct a multi-granularities feature for pseudo-label estimation. However, most of these part based methods concatenate those features directly, which increase the computing cost at the testing stage. Meanwhile, concatenating features directly will also overlook the latent non-linear relationships between different granularities in feature maps. Our method expolre the latent non-linear relationships among different features and improve the testing efficiency by encoding features to a lower dimension with a hyper-network (HN).

III. Methodology

Under the setting of unsupervised domain adaptive person re-ID, we are provided a labeled source dataset $\{X_s, Y_s\}$ that includes N_s images. Each image $x_{s,i}$ is associated with an identity $y_{s,i}$, where $y_s^i \in \{1, 2, ..., P_s\}$. P_s is the number of identities in the source training set. In addition, an unlabeled target dataset X_t that contains N_t images are also provided. Our goal is to make use of both labeled and unlabeled images to learn a model that can extract discriminative embeddings for samples of the target domain.

A. Method Overview

The overview of our noisy pseudo-label processing (NPLP) is shown in Fig.2. We first pre-train a re-ID model on the

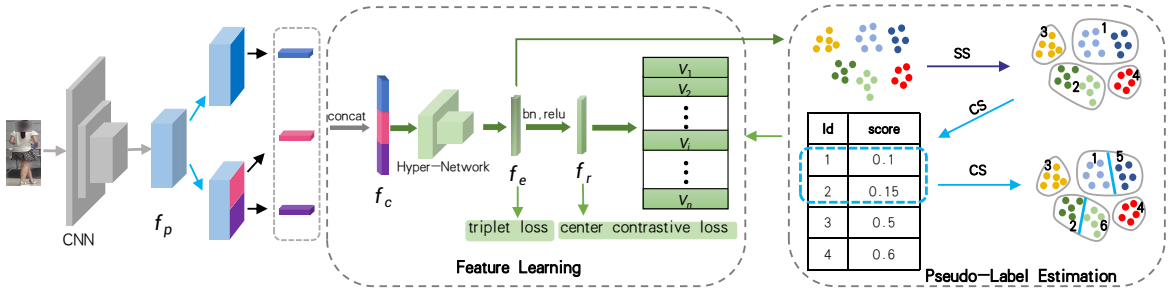

Fig. 2. Overview of our noisy pseudo-label processing approach (NPLP). **SS** means the **Startup Stage**, **CS** means the **Correcting Stage**. We first increase the recall of the clustering results at the Startup Stage to suppress the fragmented noise. At the Correcting Stage, we compute the self-assessment score to evaluate the quality of each cluster and then perform further clustering among those noisy clusters (e.g. cluster "1" and "2") to obtain higher quality pseudo-labels. For the feature learning, we introduce a hyper-network to take advantages of multi-granularities features and improve testing efficiency.

$\{X_s, Y_s\}$ in a supervised manner, and then train the pre-trained model on X_t with NPLP iteratively. Our approach contains two key steps, i.e., pseudo-label estimation and unsupervised part based feature learning. Pseudo-label estimation aims to address those two types of pseudo-label noises mentioned above and generate high quality pseudo-labels. The unsupervised part based feature learning utilizes the generated pseudo-labels to further train the re-ID model. Our method will be detailed in the following sections.

B. Supervised Pre-training

We utilize Resnet50 [29] pre-trained on ImageNet [30] as our backbone and pre-train a re-ID model $\mathcal{M}$ on source domain $\{X_s, Y_s\}$. We employed hard-batch triplet loss [31] and softmax cross-entropy loss to train our re-ID model on $\{X_s, Y_s\}$. The hard-batch triplet loss and the softmax cross-entropy loss is formulated as follows:

$$L_{triplet} = \sum_{i=1}^{P}\sum_{a=1}^{K}[\max_{p=1..K} ||x_a^i - x_p^i||_2 - \min_{\substack{n=1...K \\ j=1...P \\ j\neq i}} ||x_a^i - x_n^j||_2 + \alpha] \tag{1}$$

$$L_{softmax} = -\frac{1}{P\times K}\sum_{i=1}^{P}\sum_{a=1}^{K}\log\frac{e^{W_i^T x_a^i}}{\sum_{c=1}^{P_s} e^{W_k^T x_a^i}} \tag{2}$$

where P, K represent the number of identities and images of each identity in a mini-batch respectively. For the hard-batch triplet loss $L_{triplet}$, x_a^i, x_p^i, x_n^j represent the features f_e extracted from anchor, positive and negative samples of identity i and j respectively, while α is the margin. For the softmax cross-entropy loss $L_{softmax}$, W_k^T is the k^{th} identity's weights in the last FC layer. The total loss for supervised pre-training is formulated as:

$$L_{sp} = L_{softmax} + L_{triplet} \tag{3}$$

By minimizing L_{sp}, the re-ID model $\mathcal{M}$ can perform well on the source domain. However, the performance of $\mathcal{M}$ drops severely on the target domain due to the problem of *domain shift*. In the following subsection, we will detail our noisy pseudo-label processing appraoch (NPLP) to further train the $\mathcal{M}$ with the unlabeled data sampled from the target domain.

C. Pseudo-Label Estimation

We adopt the idea of divide and conquer to suppress the two types of noises mentioned above to generate proper pseudo-labels for re-ID model to learn. In particular, we adopt different pseudo-label estimation strategies to suppress one kind of noise at each training stages. At the *Startup Stage*, we suppress the fragmented noise and make mixed noise dominant. Then, we filter out those pseudo-labels with mixed noise and further perform pseudo-label estimation within those noisy clusters to generate pseudo-labels with higher quality at the *Correcting Stage*.

a) Startup Stage: At this stage, we aim to suppress the ratio of the fragmented noise by increasing the recall of the clustering results. We choose the most commonly used clustering algorithm, DBSCAN [17] for pseudo-label estimation. By tuning the maximum distance between neighbors, which is the most important parameter in DBSCAN [17], we are able to loose the clustering criterion and increase the probability of the same person's images captured by different cameras owning the same pseudo-label. Therefore, at this stage, the maximum distance between neighbors ϵ_e in DBSCAN [17] is computed as follows:

$$\epsilon_e = \frac{1}{\alpha N}\sum_{p=1}^{\alpha N}\mathcal{S}(d(x_{c_i,i}, x_{c_j,j})), \quad \forall x_{c_i,i}, x_{c_j,j} \in X_t, i\neq j, c_i \neq c_j \tag{4}$$

where c_i means the camera-ID of $x_{c_i,i}$ and $d(x_{c_i,i}, x_{c_j,j})$ means the feature distance across camera person image pairs. $\mathcal{S}(d(x_{c_i,i}, x_{c_j,j}))$ means sorting all $d(x_{c_i,i}, x_{c_j,j})$ from lowest to highest and N is the total number of possible pairs. We set the average value of top αN as the maximum distance between neighbors ϵ_e for pseudo-label estimation. The ϵ_e is large enough to group most similar pairs and increase the recall of the clustering results. Therefore, the mixed noise dominates and the fragmented noise is suppressed. We define the training set with mixed noise as follows:

$$X_t^m = \{C_1^{N_1}, ..., C_k^{N_k}, ..., C_c^{N_c}\} \tag{5}$$

Algorithm 1 Noisy Pseudo-Label Processing Approach

Input: labeled source dataset $\{X_s, Y_s\}$; unlabeled target dataset X_t; test set $X_{t,t}$

Train CNN model $M(x_i, f_e)$ with X_s and Y_s.
Initialize : N : total iterations , I_e, I_c : number of iterations at the startup stage and the correcting stage, E : number of epochs for each iteration.
while $iteration < N$ **do**
  Extract feature f_e for each sample in X_t.
  Generate pseudo-labels with mixed noise by constructing training set X_t^m.
  if $iteration > I_e$ **then**
    Compute the SA score of each cluster in X_t^m.
    Construct the low-quality training set $X_t^{m,l}$.
    Perform further clustering among $X_t^{m,l}$ and construct the high-quality training set X_t^c
  end if
  Construct the lookup table **V**
  Train $M(x_i, f_e, f_r)$ with L_{el} for E epochs
end while
return the CNN model $M(x_i, f_e)$

where $C_k^{N_k}$ means that the k^{th} cluster in X_t^m has N_k samples, and c is the number of clusters. We set the pseudo-label of each sample in $C_k^{N_k}$ as k. As shown in Fig.1, the appearance of images owning different pseudo-labels is significantly different and images with different pseudo-labels are easy to identify. Therefore, pseudo-labels with mixed noise can provide "easy" knowledge for $\mathcal{M}$ to learn and improve its performance on X_t at this stage.

b) Correcting Stage: As the training progresses, the re-ID model $\mathcal{M}$ can capture the partial distribution of the target domain and will fit to the mixed noise. Therefore, we change the training process into the Correcting Stage to filter out those noisy pseudo-labels and then suppress the mixed noise. At this stage, we first perform the same pseudo-label estimation method as described in Startup Stage to obtain the training set X_t^m with mixed noise. Then, we introduce the self-assesment score to evaluate the quality of each cluster and operate fine granularity clustering within those noisy clusters to generate higher quality pseudo-labels. Firstly, we formulate the intra-cluster dissimilarity a_i and inter-cluster dissimilarity b_i of $\forall x_i \in C_k^{N_k}$ as follows:

$$a_i = \frac{1}{N_k - 1} \sum_{\substack{j=1 \\ j \neq i}}^{N_k} d(x_i, x_j), \forall x_j \in C_k^{N_k} \tag{6}$$

$$b_i = \frac{1}{N_n - N_k} \sum_{o=1}^{N_n - N_k} d(x_i, x_o), \forall x_o \notin C_k^{N_k} \tag{7}$$

where $d(x_i, x_j)$ is the euclidean distance of feature f_e between x_i, x_j. For each x_i in $C_k^{N_k}$, the smaller a_i and the bigger b_i , the higher probability that x_i belongs to the cluster $C_k^{N_k}$. Taking account of both intra-cluster dissimilarity and inter-cluster dissimilarity, we can obtain the self-assessment (SA) score of each cluster in X_t^m by averaging the silhouette coefficient of samples within the cluster, which is formulated as follows:

$$S_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{b_i - a_i}{\max\{a_i, b_i\}}, \forall x_i \in C_k^{N_k} \tag{8}$$

The smaller S_k means the higher probability that $C_k^{N_k}$ is a low-purity cluster. Afterwards, we regard the $\mathcal{K}$ clusters with the lowest self-assesment score in the X_t^m contain a lot of mixed noise, and construct a set of low-quality clusters $X_t^{m,l}$. The rest of clusters compose an another high-purity cluster set $X_t^{m,h}$. Specifically, we set $\mathcal{K} = 0.1 \times c$ where c is the number of clusters in X_t^m. Therefore, we can construct a low-quality cluster set $X_t^{m,l}$. Clusters in $X_t^{m,l}$ have higher probability of containing images of different person, causing the mixed noise in pseudo-labels. Hence, we operate fine granularity clustering within each $C_k^{N_k} \in X_t^{m,l}$ to split each $C_k^{N_k}$ into several smaller and purer sub-clusters $C_{k,i}^{N_i}$. Accordingly, samples in all new sub-clusters $C_{k,i}^{N_i}$ and high-purity clusters $X_t^{m,h}$ construct a training set X_t^c with higher purity. Then the model is provided purer pseudo-labels to learn. Specifically, the maximum distance between neighbors for DBSCAN [17] based pseudo-label estimation at this stage is computed as:

$$\epsilon_c = \beta \epsilon_e \tag{9}$$

where ϵ_e is the maximum distance between neighbors used at startup stage, and the β is the attenuation factor. As a result, we obtain higher quality pseudo-labels to further train the re-ID model.

D. Feature Learning

We train the model $\mathcal{M}$ with the pseudo-labels obtained by methods described in Sec III-C. As recent part based re-ID works [1], [2], [9], [16] have shown the effectiveness of combining multi-granularities features in improving the performance of person re-ID, we adopt a simple but effective end-to-end feature learning strategy to make full use of the discriminative information among different granularities.

a) Pyramidal Average Pooling: As shown in Fig.2, we first extract the feature maps by $\mathcal{M}(I)$ and obtain the 3D tensor $\mathcal{F}$ with the size of $C \times H \times W$ for each image in X_t. Then, we slice $\mathcal{F}$ into I different granularities. In particular, for i^{th} granularity ($i \in I$), we uniformly split the $\mathcal{F}$ into i stripes, which shape are $C \times \frac{H}{i} \times W$. Then we operate global average pooling on those stripes and obtain features with various granularities. We call the above operations as pyramidal average pooling (PAP), while PAP-I means I granularities we would utilize.

b) Hyper-Network: To further improve the performance and the testing efficiency of $\mathcal{M}$ on the target domain, we propose a hyper-network (HN) to explore the potential non-linear relationships among different features and encode the concatenate feature f_c to a lower dimension but more robust feature f_e for training and testing. Our hyper-network is composed of a single fully connected (FC) layer.

TABLE I

EFFECTIVENESS OF DIFFERENT TRAINING STAGES AND DIFFERENT PAP-I DESCRIBED IN SEC. III-C AND SEC. III-D RESPECTIVELY. **DT**: DIRECT TRANSFER. **W/ SS**: TRAINING $\mathcal{M}$ WITH STARTUP STAGE ONLY. **NPLP**: TRAINING $\mathcal{M}$ WITH OUR WHOLE NPLP. NOTE THAT HYPER-NETWORK SHOWN IN FIG 2 IS USED IN ALL EXPERIMENTS.

Methods		DukeMTMC-reID → Market1501				Market1501 → DukeMTMC-reID			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
PAP-1 DT		50.4	67.4	73.8	23.5	26.8	42.4	48.8	13.8
PAP-1	w/ SS	86.8	94.2	96.4	70.6	72.1	82.6	85.6	52.5
	NPLP	88.0	94.9	96.5	73.2	77.6	86.4	89.2	61.3
PAP-2	w/ SS	88.7	94.8	96.6	71.8	75.0	84.6	87.5	56.5
	NPLP	**89.4**	**95.4**	**96.8**	**74.3**	77.8	86.6	89.4	61.5
PAP-3	w/ SS	87.2	94.5	96.4	68.4	76.8	86.8	89.4	59.6
	NPLP	88.3	94.7	96.6	69.9	78.2	86.9	89.7	60.5
PAP-4	w/ SS	87.5	94.5	96.2	68.8	76.9	87.0	89.6	59.4
	NPLP	88.7	94.6	96.6	70.0	**78.5**	**87.0**	**89.7**	**61.3**

c) Iterative Training: The proposed NPLP approach alternates between estimating pseudo-labels and training the model $\mathcal{M}$. Specifically, we assign the cluster-ID of each samples generated by methods described in Sec III-C as pseudo-label and train $\mathcal{M}$ by minimizing total intra-cluster variance and maximizing the inter-cluster variance using triplet loss described in III-B and Center Contrastive Loss (CC). The CC is formulated as:

$$L_{cc} = -\frac{1}{P \times K} \sum_{i=1}^{P} \sum_{a=1}^{K} \log \frac{e^{V_i^T x_a^i}}{\sum_{j=1}^{c} e^{V_j^T x_a^i}} \tag{10}$$

where c is the number of clusters, V_i is the centroid of $C_i^{N_i}$ which is stored in a lookup table $\mathbf{V}$. For each image x_a^i, we forward the informative feature f_e through batch normalization [32], ReLU [33], and obtain the feature $f_{r,a}^i$. Then we update the i-th cluster's centroid using the following form at the stage of backward-propagation,

$$V_i = (V_i + f_{r,a}^i)/2 \tag{11}$$

Therefore, our total loss for exploring learning on target data is formulated as follows:

$$L_{el} = L_{triplet} + L_{cc} \tag{12}$$

The re-ID model $\mathcal{M}$ and the quality of clusters can be refined mutually by updating the parameters of $\mathcal{M}$ and the pseudo-labels iteratively. The training process of our NPLP is described in algorithm. 1.

IV. EXPERIMENT

A. Datasets and Evaluation Protocol

We evaluate the proposed method on three large-scale person re-ID benchmarks : Market1501 [34], DukeMTMC-reID [35], [36], and MSMT17 [37].

a) Market1501: A large scale benchmark that contains 12,936 images of 751 identities for training and 19,732 images of 750 identities for testing.

b) DukeMTMC-reID: Another large scale benchmark that is derived from DukeMTMC [35]. It contains 16,522 images of 702 identities for training and 19,889 images of 702 identities for testing.

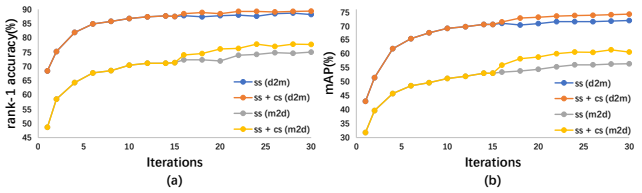

Fig. 3. Performance-iteration curve on Market1501 and DukeMTMC-reID. "SS" means training the model with Startup Stage only while "SS + CS" means training the model with Startup Stage and Correcting Stage. "d2m" means using DukeMTMC-reID as source dataset and Market1501 as target dataset while "m2d" means using Market1501 as source dataset and DukeMTMC-reID as target dataset.

c) MSMT17: The largest person re-ID benchmark which contains 126,441 images of 4,101 identities. These images are captured by 15 cameras during 4 days.

d) Evaluation Protocol: The Cumulative Matching Characteristic (CMC) curve at *Rank-1, Rank-5, Rank-10* and mean Average Precision (mAP) are used to evaluate the performance of the proposed approach.

B. Implementation Details

We first pre-train a model on the labeled source dataset following the strategy described in CamStyle [10]. For the target dataset, we set the mini-batch size as 64, where P are K are 16 and 4, respectively. Input images are resized to 256×128. Specifically, we employ random cropping, flipping, and random erasing [38] strategies for data augmentation. We use the SGD optimizer and set the learning rate as 1.2×10^{-3}. We train the model for 15 iterations at the *Startup Stage* and then change the training process into *Correcting Stage* for the rest iterations.

C. Ablation Studies

We conducted ablation studies on Market1501 [34] and DukeMTMC-reID [35], [36] to analyze the effectiveness of each component in our NPLP.

a) Effectiveness of the Startup Stage: As shown in Table I, rows with "w/ SS" are the experiment results of training re-ID model $\mathcal{M}$ in Startup Stage only. Specifically, the rank-1 accuracy and mAP of "w/ SS" are up to 38.3% and 48.3% higher than "DT"(direct transfer) respectively when $\mathcal{M}$ is

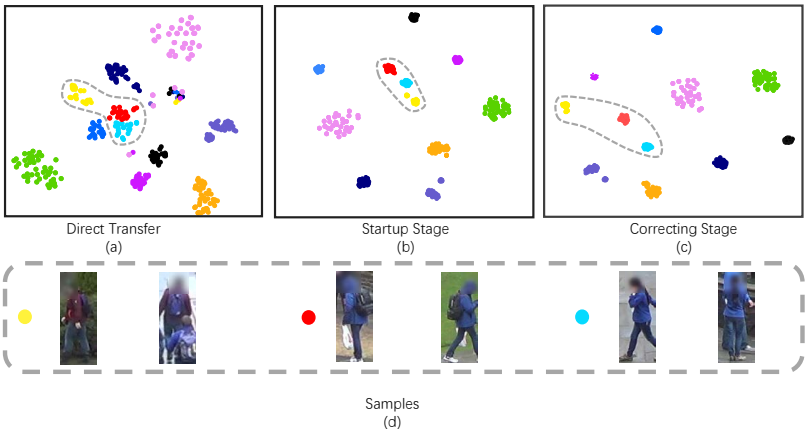

Fig. 4. T-SNE visualization of the learned feature embeddings training with different stages. In the first row, points of the same color represent images of the same identity. When we train the re-ID model with Startup Stage and Correcting Stage (subgraph (c)), the decision boundary of those similar images (points in yellow, red and blue) can be found better than Startup Stage Only (subgraph(b)) and Diret Transfer (subgraph(a)).

tested on Market1501. Similarly, the rank-1 accuracy and mAP of "w/ SS" are up to 51.1% and 45.8% higher than "DT" respectively when $\mathcal{M}$ is tested on DukeMTMC-reID. Moreover, as shown in Fig.3(a) and Fig.3(b), the performance of $\mathcal{M}$ on the target domain increase sharply in the first several iterations of Startup Stage, and then tends to rise gently. It infers that though we do not process mixed noise in pseudo-labels at the Startup Stage, the re-ID model $\mathcal{M}$ can still learn the knowledge of the target domain from the pseudo-labels and improve the quality of pseudo-labels iteratively. However, the re-ID model $\mathcal{M}$ tends to overfit to those mixed noises in pseudo-labels in the later iterations of Startup Stage, which makes its performance on target domain stops increasing.

b) Effectiveness of the Correcting Stage: As shown in Table I, rows with "NPLP" are the experiment results of training $\mathcal{M}$ in Startup Stage and Correctring Stage. The rank-1 accuracy and mAP of "NPLP" are up to 1.2% and 2.6% higher than "w/ SS" respectively when $\mathcal{M}$ is tested on Market1501. Similarly, the rank-1 accuracy and mAP of "NPLP" exceed the "w/ SS" by 0.5%-5.5% and 0.9%-8.8% when $\mathcal{M}$ is tested on DukeMTMC-reID. As shown in Fig 3(a) and Fig 3(b), when the training process is changed into Correcting Stage, the performance of $\mathcal{M}$ on target domain rises more than training $\mathcal{M}$ without Correcting Stage. It proves that our correcting stage can filter out those noisy pseudo-labels and generate purer pseudo-labels for $\mathcal{M}$ to learn and further improve the performance of the re-ID model. Moreover, as shown in Fig.4, when training with Correcting Stage, the decision boundary of those similar images can be found out, leading to the further improvement of the model's performance.

c) Effectiveness of Different PAP-I: As shown in Table I, we conducted several experiments to evaluate the effectiveness of the Noise-Reducing Strategy. PAP-i means splitting feature maps into i granularities, thus PAP-1 is the same as the global average pooling strategy. Without Correcting Stage, the rank-1 accuracy and mAP of PAP-I ($I > 1$) are up to 1.9% and 1.2% higher than PAP-1 respectively when $\mathcal{M}$ is tested on Market1501, and are up to 5.8% and 7.1% higher than PAP-1 respectively when $\mathcal{M}$ is tested on DukeMTMC-reID. Similarly, with NPLP, the rank-1 accuracy and mAP of PAP-I ($I > 1$) is up to 1.4% and 1.1% higher than PAP-1 respectively when $\mathcal{M}$ is on Market1501, and is up to 0.9% and 0.2% higher than PAP-1 respectively when $\mathcal{M}$ is tested on DukeMTMC-reID.

TABLE II
THE EFFECTIVENESS OF HYPER NETWORK DESCRIBED IN III-D. **PAP-2**: UTILIZING 2 GRANULARITIES OF FEATURE MAP FOR PYRAMIDAL AVERAGE POOLING. **DC**: CONCATENATE PAP-2 FEATURES DIRECTLY. **HN**: HYPER-NETWORK.

Methods	DukeMTMC-reID → Market1501			
	Rank-1	Rank-5	Rank-10	mAP
PAP-2 w/DC	87.1	94.9	96.4	68.7
PAP-2 w/HN	88.7	94.8	96.6	71.8
Methods	**Market1501 → DukeMTMC-reID**			
	Rank-1	Rank-5	Rank-10	mAP
PAP-2 w/DC	69.3	81.1	84.5	50.3
PAP-2 w/HN	75.0	84.6	87.5	56.5

d) Effectiveness of Hyper-Network: We also explore the effectiveness of the proposed Hyper-Network and the results are shown in Table II. The rank-1 accuracy and mAP of fusing features of different granularities with a Hyper-Network are 2.3% and 5.6% higher than directly concatenating multi-granularity features when $\mathcal{M}$ is tested on Market1501. Similarly, the rank-1 accuracy and mAP of fusing features of different granularities with a Hyper-Network are 8.5% and 11.2% higher than directly concatenating multi-granularity features when $\mathcal{M}$ is tested on DukeMTMC-reID. This is because our Hyper-Network can effectively explore the latent relationships among features of different granularities and learn a more discriminative feature.

D. Comparison with State-of-the-art Methods

We compare the performance of our approach with state-of-the-art unsupervised person re-ID methods on Market1501, DukeMTMC-reID and MSMT17 in Table.III and Table.IV.

TABLE III

COMPARISON OF OUR NPLP WITH STATE-OF-THE-ARTS UNSUPERVISED DOMAIN ADAPTIVE PERSON RE-ID METHODS ON MARKET1501 AND DUKEMTMC-REID. **BOLD** INDICATES THE BEST AND UNDERLINED THE RUNNER-UP. ∗ MEANS THE PSEUDO-LABEL BASED METHODS.

Methods	Venue	DukeMTMC-reID → Market1501				Market1501 → DukeMTMC-reID			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
LOMO [39]	CVPR15	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
BOW [34]	ICCV15	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
TJ-AIDL [40]	ECCV18	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
CamStyle [10]	CVPR18	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
PAUL [14]	CVPR19	66.7	-	-	36.8	56.1	-	-	35.7
ECN [12]	CVPR19	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
UDA∗ [15]	PR20	75.8	89.5	93.2	53.7	68.4	80.1	83.5	49.0
PAST∗ [16]	ICCV19	78.4	-	-	54.6	72.4	-	-	54.3
SSG∗ [9]	ICCV19	80.0	90.0	92.4	58.3	73.0	80.6	83.2	53.4
MMCL∗ [41]	CVPR20	84.4	92.8	95.0	60.4	72.4	82.9	85.0	51.4
AD-Cluster∗ [42]	CVPR20	86.7	94.4	96.5	68.3	72.6	82.5	85.5	54.1
MMT∗ [27]	ICLR20	87.7	94.9	96.9	71.2	78.0	**88.8**	**92.5**	**65.1**
Ours (NPLP)	-	**89.4**	**95.4**	**96.8**	**74.3**	**78.5**	87.0	89.7	61.3

TABLE IV

COMPARISON OF THE PROPOSED APPROACH WITH STATE-OF-THE-ART METHODS ON MSMT17. **BOLD** INDICATES THE BEST AND UNDERLINED THE RUNNER-UP.

Methods	Market1501 → MSMT17			
	Rank-1	Rank-5	Rank-10	mAP
PTGAN [37]	10.2	-	24.4	2.9
SSG [9]	31.6	-	49.6	13.2
MMCL [41]	40.8	51.8	56.7	15.1
MMT [27]	49.2	63.1	68.8	22.9
Ours (NPLP)	**52.8**	**64.3**	**69.5**	**23.3**
Methods	DukeMTMC-reID → MSMT17			
	Rank-1	Rank-5	Rank-10	mAP
PTGAN [37]	11.8	-	27.4	3.3
SSG [9]	32.2	-	51.2	13.3
MMCL [41]	43.6	54.3	58.9	16.2
MMT [27]	50.1	63.9	**69.8**	23.3
Ours (NPLP)	**52.6**	**64.9**	69.6	**23.4**

We achieve 89.4% rank-1 accuracy and 74.3% mAP on Market1501 [34], which exceed the pseudo-label estimation based methods UDA [15], SSG [9], PAST [16], MMCL [41], AD-Cluster [42] by 2.7%-20% and 6%-36% respectively. The rank-1 accuracy of NPLP is 78.5% when tested on DukeMTMC-reID [35], [36]. This is higher than all the other methods in Table.III. It is worth noting that the SSG [9] utilizes the multi-granularity features independently and the PAST [16] concatenate the multi-granularity features directly, which can not explore the abundant features and robust correlations between samples effectively. Moreover, our method outperforms the best published method MMT [27] when tested on Market1501 and MSMT17. MMT [27] needs auxiliary models for training, which increases the training costs. Therefore, our method is shown to compare favourably with the state-of-the-art methods.

E. Parameter Analysis

In this section, we analyze the sensitivities of our approach to two important hyper-parameters α and β that are described in Eq(4) and Eq(9) respectively. The experiment results are shown in Fig. 5.

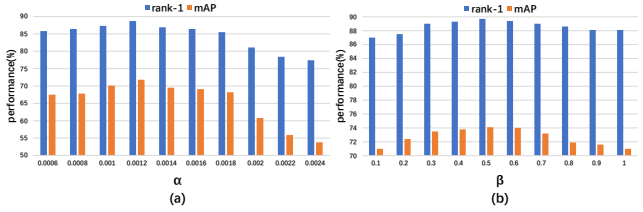

Fig. 5. The sensitivity of NPLP to α and β that are described in Eq(4) and Eq(9) respectively. We used DukeMTMC-reID as source domain and Market1501 as target domain.

a) Parameter α at Startup Stage: In Fig. 5(a), we investigate the effect of the parameter α at Startup Stage Using a lower α leads to a lower ϵ_e for pseudo-label estimation at the Startup Stage. Therefore, similar pairs may be grouped into different clusters, generating more fragmented noise in pseudo-labels. On the other hand, using a higher α leads to a higher ϵ_e for pseudo-label estimation at the Startup Stage, generating more mixed noise in pseudo-labels. Due to the large scale of our datasets and a large number of the cross-camera image pairs, a small change of α has a discernible impact on the final performance. The best results are produced when α is around 0.0012.

b) Parameter β at Correcting Stage: In Fig. 5(b) , we compare the effect of different β in Eq(9). Using a lower β leads to a lower ϵ_c for further pseudo-label estimation at the Correcting Stage. Thus, the standard of different images being grouped into the same cluster will be tighter, leading to a high probability that similar pairs being assigned different pseudo-labels. Therefore, fragmented noise in pseudo-labels will increase. Meanwhile, a higher β produces a higher ϵ_c, leading to a looser standard of different images being grouped into the same cluster. When $\beta = 1$, our method reduces to training the re-ID model for Startup Stage only. Therefore, higher β can not handle mixed noise effectively. The best results are produced when β is around 0.5.

V. Conclusion

In this paper, we analyze two types of pseudo-label noises and their causes in pseudo-label based unsupervised person re-ID methods. Then, we propose a Noisy Pseudo-Label Processing (NPLP) approach to relieve these two kinds of noises base on the idea of divide and conquer. We group those similar images to suppress the fragmented noise and train the re-ID model with pseudo-labels that contain mixed noise at the Startup Stage. Then, we evaluate the quality of the pseudo labels and then purify those noisy pseudo-labels at the Correcting Stage and further improve the performance of the re-ID model. Furthermore, we build a feature learning strategy for unsupervised re-ID task and learn from the denoised pseudo-labels iteratively. Extensive experiments on three benchmark show that our noisy pseudo-label processing approach significantly outperforms state-of-the-art unsupervised re-ID models by clear margins.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61876065), the Natural Science Foundation of Guangdong Province (2018A0303130022), the Science and Technology Program of Guangzhou (201904010200), the Science and Technology Planning Project of Guangdong Province, China (No. 2016A010101012), the Guangzhou Science and Technology Program key projects (202007040002), and the China Southern Power Grid (Grant No. GDKJXM20185761, GDKJXM20200484).

References

[1] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline)," in *ECCV*, 2018.

[2] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *CVPR*, 2019.

[3] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *ACM MM*, 2018.

[4] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI*, 2019.

[5] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *ICCV*, 2017.

[6] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *CVPR*, 2018.

[7] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018.

[8] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *CVPR*, 2018.

[9] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *ICCV*, 2019.

[10] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *CVPR*, 2018.

[11] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *ECCV*, 2018.

[12] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *CVPR*, 2019.

[13] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *CVPR*, 2018.

[14] Q. Yang, H. Yu, A. Wu, and W. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *CVPR*, 2019.

[15] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit*, 2020.

[16] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *(ICCV)*, October 2019.

[17] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD*, 1996.

[18] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *CVPR*, 2017.

[19] A. Vahdat, "Toward robustness against label noise in training deep discriminative neural networks," in *NIPS*, 2017.

[20] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks." in *AAAI*, 2017.

[21] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *NIPS*, 2018.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[23] K. Lee, X. He, L. Zhang, and L. Yang, "Cleannet: Transfer learning for scalable image classifier training with label noise," in *CVPR*, 2018.

[24] J. Huang, L. Qu, R. Jia, and B. Zhao, "O2u-net: A simple noisy label detection approach for deep neural networks," in *ICCV*, 2019.

[25] Y. Kim, J. Yim, J. Yun, and J. Kim, "NLNL: negative learning for noisy labels," in *ICCV*, 2019.

[26] T. Yu, D. Li, Y. Yang, T. M. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," 2019.

[27] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *ICLR*, 2020.

[28] H. Huang, W. Yang, X. Chen, X. Zhao, K. Huang, J. Lin, G. Huang, and D. Du, "Eanet: Enhancing alignment for cross-domain person re-identification," *CoRR*, 2018.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[31] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, F. R. Bach and D. M. Blei, Eds., 2015.

[33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, J. Fürnkranz and T. Joachims, Eds., 2010.

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[35] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*, 2016.

[36] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017.

[37] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *CVPR*, 2018.

[38] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[39] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[40] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *ECCV*, 2018.

[41] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *CVPR*, 2020.

[42] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, and Y. Tian, "Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification," in *CVPR*, 2020.