NPLP: A Noisy Pseudo Label Processing Approach for Unsupervised Cross-Domain Person Re-Identification

ABSTRACT

The performance of person re-identification (re-ID) has achieved great improvement with massive labeled data but still remains poor under the unsupervised cross-domain environment. Most of the existing methods utilize pseudo label estimation to cast the unsupervised cross-domain person re-ID into a supervised problem, whose performance is highly dependent on the quality of pseudo labels. The noisy pseudo labels may lead to the collapse of the re-ID model. In this paper, we propose a noisy pseudo label processing framework (NPLP) to reduce the noise of pseudo labels and learn from the noisy pseudo labels iteratively. In particular, NPLP contains two strategies, namely, noise-reducing strategy and self-correcting strategy. Specifically, the noise-reducing strategy integrates features in coarse-to-fine granularities for robust pseudo label estimation, which can provide high quality supervised signals to improve the re-ID model. On the other hand, the self-correcting strategy first trains the re-ID model with noisy pseudo labels iteratively. Then it evaluates the quality of the pseudo labels and corrects those low-quality pseudo labels to generate more reliable pseudo labels for the re-ID model to learn. By running the NPLP framework iteratively, noisy pseudo labels can be gradually purified, and thus the re-ID model can be optimized with more reliable pseudo labels. Extensive evaluations on three benchmarks show that our noisy pseudo label processing framework significantly outperforms state-of-the-art unsupervised re-ID models by a clear margin.

KEYWORDS

Person Re-Identification, Unsupervised Domain Adaptation, Noisy Label Processing

ACM Reference Format:

. 2018. NPLP: A Noisy Pseudo Label Processing Approach for Unsupervised Cross-Domain Person Re-Identification. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Person re-identification aims to retrieve the target person from the surveillance video. Recently, most person re-ID algorithms[23, 26, 34] learn with manually labeled data and have achieved impressive performance. However, obtaining labeled data for person re-ID is quite time-consuming and expensive, so that such supervised methods have limited scalability and usability in real-world applications. How to effectively learn a discriminative model on massive unlabeled data has been a challenging problem. Some methods[15, 16, 27] have been proposed to tackle the unsupervised person re-ID problem. 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[2, 4, 5, 17, 22, 30, 32, 38–40] are proposed to make use of both labeled data(from source domain) and unlabeled data(from target domain) during training and improve the re-ID model's performance on the target domain.

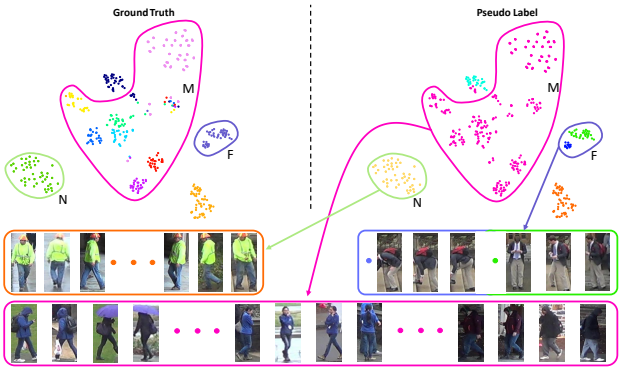

Figure 1: T-SNE visualization of the initial feature embeddings on a part of DukeMTMC-reID. Points of the same color in the left represent images of the same identity and represent images that are assigned the same pseudo label in the right. There are mainly two types of noise caused by pseudo-label estimation based methods: mixed noise (the circle M)and fragmented noise (the circle F).

The majority of the above unsupervised cross-domain methods[4, 5, 17, 22, 32] are based on pseudo label estimation, which extract features of unlabeled data using the pre-trained model and assigned each sample a pseudo label using unsupervised cluster methods (e.g., k-means and DBSCAN[3]). However, their performance is highly dependent on the quality of pseudo labels, and the noisy pseudo labels may lead to the collapse of the re-ID model. Unfortunately, there is seldom pseudo label estimation based method consider the effect of pseudo label noise to the best of our knowledge.

As shown in Fig 1, there are mainly two types of pseudo label noise produced by the label estimation procedure, as defined as mixed noise and fragmented noise in this paper, 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. The distribution of these two types of noise varies in different pseudo label estimation methods.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

In this paper, we propose a noisy pseudo labels processing framework (NPLP) to reduce the noise of pseudo labels and learn from the noisy labels iteratively. In particular, we propose a **Noise-Reducing Strategy** to integrate features in coarse-to-fine granularities with a hyper-network to construct a robust feature for pseudo label estimation, so as to reduce the noisy pseudo labels significantly.

Furthermore, a **Self-Correcting Strategy** is introduced by us to handle the two types of pseudo label noise discussed above in two key stages. At the first stage, we group those similar pairs base on the consideration of cross-camera pairs' distance and increase the probability of the same person's images captured by different cameras owning the same pseudo label. Therefore, the fragmented noise can be suppressed in this way, and the mixed noise dominates. As the network trends to first learn the knowledge from the samples, which are "easy" to fit[1], the performance of the re-ID model on the target domain can be improved quickly. Therefore we call the first stage as Eager Stage.

Then in the second stage, we propose a self-assessment (SA) score to evaluate each cluster's quality and find out the noisy clusters. We then further perform pseudo label estimation within those noisy clusters with lower SA score to correct the pseudo labels. We name the second stage as the Correcting Stage because we filter out images that are difficult for the re-ID model to identified and try to assign them a higher quality pseudo label than the Eager Stage. The experiment shows that the performance of the re-ID model boosts significantly with these two stages running iteratively.
Our contributions are as follows:

- We propose the problem of handling pseudo label noise for pseudo label estimation based unsupervised cross-domain person re-ID methods, and two main noises, namely, mixed noise and fragmented noise are identified.
- A noisy pseudo label processing framework (NPLP) is proposed to reduce pseudo label noise by constructing a robust feature and address those two types of noise with a self-correcting strategy.
- Comprehensive experiments are conducted on three large-scale benchmarks, and the results demonstrate that our simple and effective method advantages the art unsupervised person re-ID to a new level.

2 RELATED WORK

Part based Person re-ID. Recent works[23, 26, 34] have shown the effectiveness of using multi-scale features to construct robust features for supervised person re-ID. Similarly, [5, 9, 30, 32] utilized multi-scale features to tackle unsupervised person re-ID. [30] leverages similarity between patches to exploit the part affinity between instances. [5] estimated pseudo label for each part independently and trained the re-ID model with the obtained pseudo labels. EANet[9] proposed Part Aligned Pooling and Part Segmentation to enhance feature alignment. [32] adopt a similar strategy to [23] to construct a multi-scale feature for pseudo label estimation. However, most of these part based methods overlook the latent non-linear relationships between different granularities in feature maps. Our method split feature map into different granularities and integrates them with a hyper-network to further explore the latent relationships among different features. Therefore, we obtain a more robust feature for pseudo label estimation.

Pseudo Label Based Unsupervised Person re-ID. Recently, many methods are proposed to tackle unsupervised person re-ID in a self-training manner[4, 5, 15, 22, 32]. They repeated two steps, assigning pseudo labels to training samples and training re-ID model with pseudo labels until convergence. Method[4] performed clustering on the unlabeled dataset and select samples that were close to the cluster centroids to train the re-ID model. SSG[5] split feature map into different parts and estimated pseudo labels for each part independently and then trained the re-ID model iteratively. Meanwhile, BUC[15] proposed a bottom-up clustering method to group similar samples into the same identity iteratively. PAST[32] consists of conservative stage and promoting stage, while the former aimed to improve feature representations, and the latter aimed 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.

Deep Learning with Noisy Label. The problem of label noise has been widely studied in deep learning. Some methods[19, 25] estimated a noise transition matrix to correct the noisy labels. Some methods[6, 24, 33] designed noise-robust loss to learn a noise-robust model. Literatures[10, 12, 13] tried to filter out noisy labels by training iteratively and then learned with the clean label. To the best of our knowledge, few methods have been proposed to address the problem of label noise in re-ID. DistributionNet[31] modeled feature uncertainty by adding extra embedding network and designed a loss to allocate uncertainty across training samples unevenly. BUC[15] made assumptions about the number of images of each identity and designed a diversity regularization to avoid pseudo label noise to some extent. Our method analyzes two types of pseudo label noise in pseudo label based methods, and designed a simple but effective training strategy to learn with noisy pseudo labels.

3 METHODOLOGY

Problem Definition. Under the setting of [38], we are provided a labeled source set $\{X_s, Y_s\}$ 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 discriminative embeddings for the target data.

3.1 Framework Overview

The overview of our noisy pseudo label processing (NPLP) framework is shown in Fig 2. We first pre-train a re-ID model on the $\{X_s, Y_s\}$ in a supervised manner, and then train the pre-trained model on X_t with NPLP framework iteratively. Our framework contains two strategies, namely, **Noise-Reducing Strategy** and **Self-Correcting Strategy**. The noise-reducing strategy constructs a robust feature by integrating features in coarse-to-fine granularities and facilitate the production of higher quality pseudo labels. The self-correcting strategy suppresses the fragmented noise and train the re-ID model with mixed noise at the eager stage and then

evaluate the quality of the generated pseudo labels and correct those pseudo labels with a lot of mixed noise at the correcting stage. The next sections will detail our method.

3.2 Supervised Pre-training

We first utilize ResNet50[7] as our backbone to pre-train a re-ID model $\mathcal{M}$ on source domain $\{X_s, Y_s\}$. Specifically, we discard the last fully connected (FC) layer and add two additional FC layers. We employed softmax cross-entropy loss with last FC layer and hard-batch triplet loss[8] with penultimate FC layer to train our re-ID model on $\{X_s, Y_s\}$, which can be formulated as:

$$L_{softmax} = -\frac{1}{P \times K} \sum_{i=1}^{P} \sum_{a=1}^{K} \log \frac{\exp^{(W_i^T x_a^i)}}{\sum_{c=1}^{P_s} \exp^{(W_c^T x_a^i)}} \tag{1}$$

$$L_{triplet} = \sum_{i=1}^{P} \sum_{a=1}^{K} [\alpha + \max_{p=1..K} ||x_a^i - x_p^i||_2 - \min_{\substack{n=1...K \\ j=1...P \\ j \neq p}} ||x_a^i - x_n^j||_2] \tag{2}$$

where P, K represent the number of identities and images of each identity in a mini-batch respectively. For the triplet loss , x_a^i, x_p^i, x_n^j represent the features f_e (shown in Fig 2) extracted from anchor, positive and negative samples of identity i and j respectively, while α is the margin. For the softmax cross-entropy loss, W_c^T is the c^{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 data. However, the performance of $\mathcal{M}$ drops severely on the target domain due to the problem of *domain shift*. The pseudo label estimation based methods[4, 5, 15, 17, 32] have proven their effectiveness on unsupervised cross-domain person re-ID. However, most of them do not consider pseudo label noise. In the following subsection, we will detail our noisy pseudo label processing framework (NPLP) to reduce pseudo label noise by extract robust features from the unlabeled data and then evaluate the quality of the pseudo labels and train the re-ID model with purified pseudo label iteratively.

3.3 Noise Reducing Strategy

Recent part based re-ID works[5, 23, 32, 34] have shown the effectiveness of utilizing multi-granularities feature maps to construct a robust feature. When the extracted features are more robust, the distance of the features of the same identity will be closer, so the images of the same identity will be more likely to be assigned the same pseudo label. Therefore, the noise in pseudo labels will be reduced. As shown in Fig 2, our noise-reducing strategy first obtains features f_p in different granularities by pyramidal average pooling(PAP) and then integrates those features using a Hyper-Network(HN) to construct a more robust feature.

Pyramidal Average Pooling. As shown in fig 2, the Pyramidal Average Pooling operation contains two steps: Pyramidal Feature Map Segmentation (PFMS) and Global Average Pooling (GAP). For each image in $I \in X_t$, we extract the feature maps by $\mathcal{M}(I)$ and obtain the 3D tensor $\mathcal{F}$ with the size of $C \times H \times W$. PFMS means slicing $\mathcal{F}$ into i different granularities, where i refers to the number of granularities. In particular, for i^{th} granularity, we uniformly split the $\mathcal{F}$ into i stripes, which shape are $C \times \frac{H}{i} \times W$. In this way, the set of i different granularities $\mathcal{G}$ owns various granularities, which can represent the images more robustly. Then we operate GAP on each stripe to obtain coarse-to-fine features f_p. Hence, we name the above operation as pyramidal average pooling (PAP), while PAP-i means i granularities we would utilize.

Hyper-Network. Concatenating different granularities features f_p directly is the comment sense of utilizing multi-granularity features for person re-ID [5, 32]. However, it may overlook the latent relationships among different channels of different features. As shown in Fig 2, we adopt a Hyper-Network to explore the potential relationships between 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. In this way, the latent non-linear relationships between different granularity features can be found, and the useless information can be filtered. Therefore, we construct a more robust feature for pseudo label estimation.

3.4 Self-Correcting Strategy

Though we construct a more robust feature for pseudo label estimation by utilizing noise-reducing strategy, it is difficult to avoid pseudo label noise by directly using the features extracted from $\mathcal{M}$ for X_t to estimate pseudo labels due to the weak generalization ability of the pre-trained re-ID model $\mathcal{M}$. Therefore, we adopt the self-correcting strategy to alleviate this problem. Our self-correcting strategy consists of two stages, namely, **Eager Stage** and **Correcting Stage**. We first suppress fragmented noise and learn with mixed noise iteratively at the eager stage. The re-ID model $\mathcal{M}$ will overfit to the mixed noise as the training processes. Then we change the training process into the correcting stage to filter out those noisy pseudo labels and generate higher quality pseudo labels for $\mathcal{M}$ to learn.

Eager Stage. As mentioned above, the intra-domain image style variations are the main cause of fragmented noise and mixed noise. To flexibly control the ratio of different noise, we choose the most commonly used clustering algorithm, DBSCAN[3] for pseudo label estimation. We set the threshold T_e of DBSCAN as follows:

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

where c_i means the camera of $x_{c_i,i}$ and $d(x_{c_i,i}, x_{c_j,j})$ means the cross-camera pairs' distance, N is the total number of possible pairs and $\mathcal{S}(d(x_{c_i,i}, x_{c_j,j}))$ means sorting all $d(x_{c_i,i}, x_{c_j,j})$ from lowest to highest. We set the average value of top αN as the threshold T_e for pseudo label estimation. Therefore, the probability of the same person's images in different cameras being assigned the same pseudo label will be higher. Most of the similar pairs will be grouped so that mixed noise dominates and the fragmented noise is suppressed. Therefore, we obtain the training set with mixed noise that is defined as follows:

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

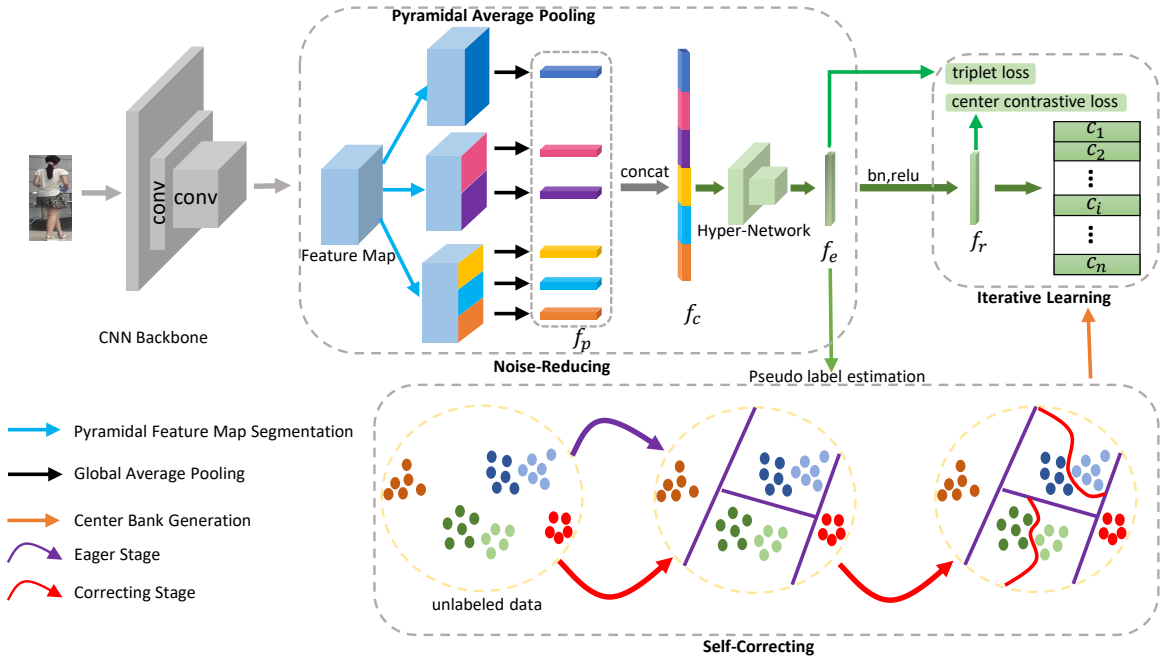

Figure 2: Overview of our noisy pseudo label processing framework (NPLP). We first operate pyramidal average pooling upon feature maps and then integrate different granularities in feature with a Hyper-Network to construct a robust feature for pseudo label estimation. Then, we group those similar pairs using a looser standard and generate pseudo label with mixed noise for re-ID model to learn at the Eager Stage. Then, the training process changes into Correcting Stage. We filter out clusters that have higher probability of containing mixed noise and operate further pseudo label estimation among them with strict standard. Thus, we obtain the purer pseudo label for re-ID model to learn.

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, despite the mixed noise, the appearance of images owning different pseudo labels is significantly different and images with different pseudo labels are easy to distinguish. Therefore, pseudo label with mixed noise can provide "easy" knowledge for $\mathcal{M}$ to learn and improve its performance on X_t at this stage.

Correcting Stage. As the training progresses, the re-ID model $\mathcal{M}$ can capture the partial distribution of the target domain but will overfit to those mixed noises. Therefore, we change the training process into the Correcting Stage to filter out those noisy pseudo labels and then handle the mixed noise. At the correcting stage, we first perform the same pseudo label estimation method as described in Eager Stage to obtain the training set X_t^m with mixed noise. Then, two steps, namely, **Noisy Pseudo Label Filtering** and **Noisy Pseudo Label Correcting** are performed after every pseudo label estimation to assign each sample in X_t^m a more reliable pseudo label, and thus further improve the performance of $\mathcal{M}$ on X_t.

(1) Noisy Pseudo Label Filtering. We first introduce the intra-cluster dissimilarity a_i and inter-cluster dissimilarity b_i of $\forall x_i \in C_k^{N_k}$ as :

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

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

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 calculate the self-assessment 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,

Algorithm 1 The Noisy Pseudo Label Processing Framework

Input: labeled source dataset $\{X_s, Y_s\}$; unlabeled target dataset X_t; unlabeled 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 : number of iterations at the eager stage, I_c : number of itreations at the correcting stage, E : number of epochs for each iteration.
while $iteration < N$ **do**
  extract robust feature f_e for each sample in X_t.
  if $iteration < I_e$ **then**
    generate pseudo labels with mixed noise by constructing X_t^m
    Construct the lookup table **V** with current pseudo labels.
    Train $M(x_i, f_e, f_r)$ on X_t^m using L_{el} for E epochs
  else
    Handle mixed noise by self-correcting : $X_t^m \rightarrow X_t^c$.
    Construct the lookup table **V** with higher quality pseudo labels.
    Train $M(x_i, f_e, f_r)$ on X_t^c using L_{el} for E epochs
  end if
end while
return A CNN model $M(x_i, f_e)$ performs well on $X_{t,t}$

we filter out the noisy pseudo labels by constructing the low-quality cluster set $X_t^{m,l}$.

(2) Noisy Pseudo Label Correcting. As mentioned before, clusters in $X_t^{m,l}$ has 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 thresold T_c for DBSCAN[3] based pseudo label estimation at this stage is set to be:

$$T_c = \beta T_e \tag{9}$$

where T_e is the threshold used at Eager Stage, and the β is the threshold attenuation factor. As a result, we obtain higher quality pseudo labels to further train the re-ID model.

3.5 Iterative Learning

The proposed Noisy Pseudo Label Processing framework trains the re-ID model $\mathcal{M}$ iteratively. In particular, we assign the cluster-ID as the pseudo label of each sample and train the network $\mathcal{M}$ by minimizing total intra-cluster variance and maximizing the inter-cluster variance using Center Contrastive Loss, which is formulated as:

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

where c is the number of clusters, V_i is the center vector of $C_i^{N_i}$ which is stored in a lookup table **V**. For each image x_a^i, we forward the informative feature f_e through batch normalization[11], ReLU[18], and obtain the feature $f_{r,a}^i$. Then we update the i-th cluster's center vector using the following form at the stage of backward-propagation,

$$V_i = mV_i + (1-m) f_{r,a}^i \tag{11}$$

where $m \in [0, 1)$ is a hyper-parameter that controls the updating rate. Therefore, our total loss for exploring learning on target data is formulated as follows:

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

After optimized by the L_{el} for several epochs, the re-ID model $\mathcal{M}$ can extract a robust feature for each sample in X_t, facilitating the generation of high-quality pseudo labels. Then the high-quality pseudo labels enable the re-ID model $\mathcal{M}$ to further discover the distribution of X_t. Therefore, 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 noisy pseudo label processing framework is described in Algorithm 1.

4 EXPERIMENT

4.1 Datasets and Evaluation Protocol

We evaluate the proposed method on three large-scale person re-identification benchmarks : Market1501[35], DukeMTMC-reID[21, 36], and MSMT17[29]. We adopt the Cumulative Matching Characteristic(CMC) curve at *Rank-1, Rank-5, Rank-10* and mean Average Precision(mAP) to evaluate the performance of the proposed approach.

4.2 Implementation Details

We first pre-train a model on the labeled source dataset following the strategy described in CamStyle[40]. For the target dataset, we set the mini-batch size as 64, where P are K are 16 and 8, respectively. Input images are resized to 256×128. Specifically, we employ random cropping, flipping, and random erasing[37] 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 **Eager Stage** and then change the training process into **Correcting Stage** for 15 iterations.

4.3 Ablation Studies

We conducted ablation studies on Market1501 and DukeMTMC-reID to analyze the effectiveness of each component in our NPLP framework.

Effectiveness of Noise-Reducing Strategy

Effectiveness of PAP-i. As shown in Table 1, 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. "w/ ES" means training re-ID model $\mathcal{M}$ without Correcting Stage and "NPLP" means training $\mathcal{M}$ with the whole framework. 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

Table 1: Effectiveness of different PAP-i and different training stages described in Sec 3.3 and Sec 3.4 respectively. SL: supervised learning, DT: direct transfer. w/ ES: training $\mathcal{M}$ with Eager Stage only. NPLP: training $\mathcal{M}$ with our whole NPLP framework. 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
SL		92.5	97.5	98.4	80.8	82.6	92.3	94.4	70.5
PAP-1 DT		50.4	67.4	73.8	23.5	26.8	42.4	48.8	13.8
PAP-1	w/ ES	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/ ES	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/ ES	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/ ES	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**
PAP-5	w/ ES	86.4	94.2	96.0	65.8	77.9	86.4	89.1	58.9
	NPLP	88.1	94.4	96.1	67.7	78.4	86.6	89.6	59.9

Table 2: the effectiveness of Hyper Network described in 3.3. PAP-2: utilizing 2 granularities of feature map for pyramidal average pooling. DC: concatenate PAP-2 features directly. HN: Hyper Networks.

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	89.4	95.4	96.8	74.3
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	77.8	86.6	89.4	61.5

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. It infers that the proposed Noise-Reducing Strategy utilizes information from feature maps of different granularities and provides a robust feature for pseudo label estimation and pseudo labels generated by this way contain less noise. However, training with the overmuch multi-granularity features may lead to a bad performance than PAP-1 as shown in Table 1. It is because that clustering and training with the overmuch scale feature may lead to the overfitting on noisy-cluster of the network and thus limit the generalization ability of the re-ID model $\mathcal{M}$.

Effectiveness of Hyper-Network. Specifically, we also explore the effectiveness of the proposed Hyper-Network and the results are shown in Table 2. 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 construct a more robust feature and reduce the probability of the re-ID model $\mathcal{M}$ overfitting to invalid information.

Effectiveness of Self-Correcting Strategy

Effectiveness of the Eager Stage. As mentioned above, rows with "w/ ES" in table 1 are the experiment results of training re-ID model $\mathcal{M}$ in Eager Stage only. Specifically, the rank-1 accuracy and mAP of "w/ ES" are up to 38.3% and 48.3% higher than "DT"(direct transfer) respectively when $\mathcal{M}$ is tested on Market1501. Similarly, the rank-1 accuracy and mAP of "w/ ES" 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 5(a) and fig 5(b), the performance of $\mathcal{M}$ on the target domain increase sharply in the first several iterations of Eager Stage, and then tends to rise gently. It infers that though we do not process mixed noise in pseudo labels at the Eager 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 Eager Stage, which makes its performance on target domain stops increasing.

Effectiveness of the Correcting Stage. As shown in table 1, rows with "NPLP" are the experiment results of training $\mathcal{M}$ in Eager Stage and Correctring Stage. The rank-1 accuracy and mAP of "NPLP" are up to 1.2% and 2.6% higher than "w/ ES" respectively when $\mathcal{M}$ is tested on Market1501. Similarly, the rank-1 accuracy and mAP of "NPLP" exceed the "w/ ES" by 0.5%-5.5% and 0.9%-8.8% when $\mathcal{M}$ is tested on DukeMTMC-reID. As shown in fig 5(a) and fig 5(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 3, when training with Correcting Stage, the decision boundary of those similar images

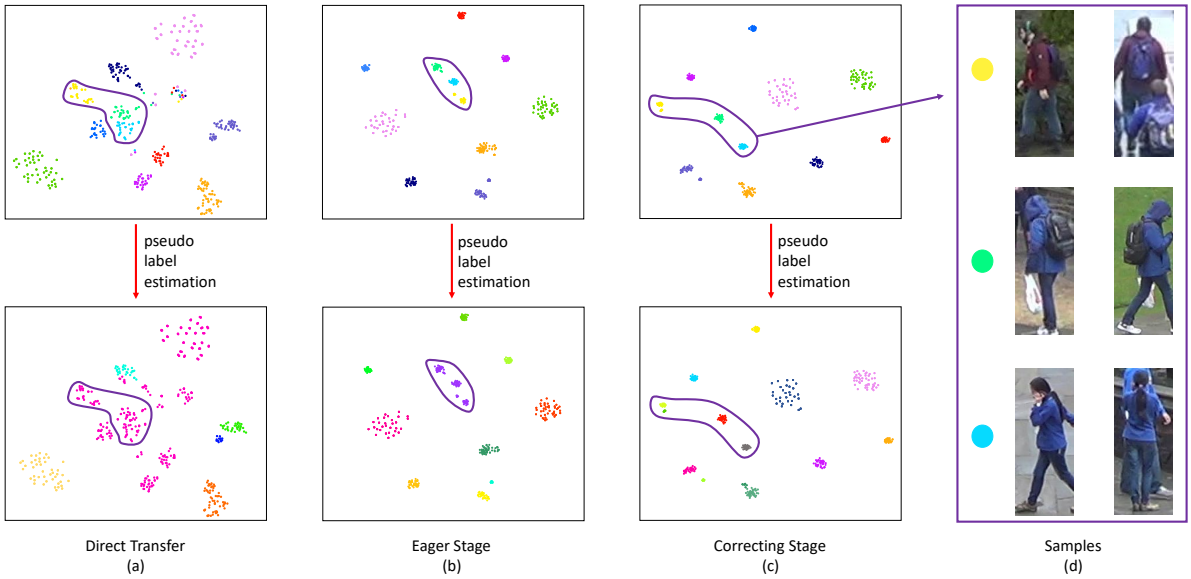

Figure 3: T-SNE visualization of the learned feature embeddings training with different stages. In the first row, points of the same color represent images of the same identity. In the second row, points of the same color represent images of the same pseudo label. When we train $\mathcal{M}$ with Eager Stage only(subgraph (b)), similar images(points in yellow, cyan, blue in the first row) will be assigned the same pseudo label, leading to mixed noise. Meanwhile, when training the re-ID model with Eager Stage and Correcting Stage(subgraph (c)), the decision boundary of those similar images can be found well.

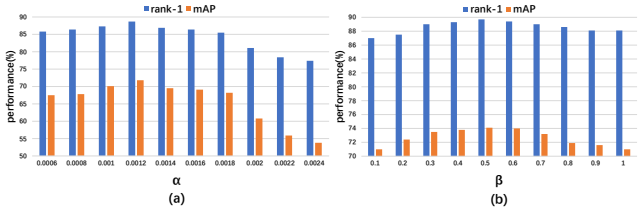

Figure 4: The sensitivity of NPLP to α and β that are described in Eq(4) and Eq(9) respectively. We used DukeMTMC-reID as source domain and Market1501 as target domain.

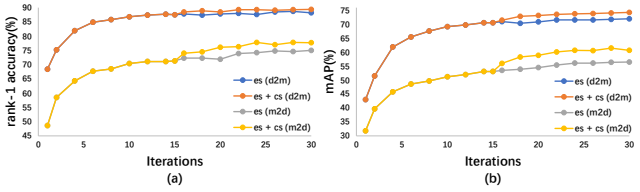

Figure 5: performance-iteration curve on Market1501 and DukeMTMC-reID for the comparison of Eager Stage (ES) and Correcting Stage (CS). d2m means using DukeMTMC-reID as source dataset and Market1501 as target dataset, and vice versa.

can be found out, leading to the further improvement of the model's performance.

4.4 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 4.

Parameter α at Eager Stage. In fig 4(a), we investigate the effect of the parameter α at Eager Stage Using a lower α leads to a lower T_e for pseudo label estimation at the Eager 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 T_e for pseudo label estimation at the Eager 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.

Parameter β at Correcting Stage. In fig 4(b) , we compare the effect of different β in Eq(9). Using a lower β leads to a lower T_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 T_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 Eager Stage only. Therefore, higher β can not handle mixed noise effectively. The best results are produced when β is around 0.5.

Table 3: Comparison of our NPLP with state-of-the-arts unsupervised domain adaptive person re-ID methods on Market1501 and DukeMTMC-reID. ∗ denotes methods based on unsupervised clustering and • means no labels are used in the methods.

Methods	Venue	Duke → Market				Market → Duke			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
LOMO[14]	CVPR15	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
Bow[35]	ICCV15	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
UMDL[20]	CVPR16	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
SPGAN[2]	CVPR18	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
TJ-AIDL[28]	ECCV18	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
CamStyle[40]	CVPR18	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
PAUL[30]	CVPR19	66.7	-	-	36.8	56.1	-	-	35.7
ECN[39]	CVPR19	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
BUC∗•[15]	AAAI19	66.2	79.6	84.5	38.3	47.4	62.6	68.4	27.5
UDA∗[22]	-	75.8	89.5	93.2	53.7	68.4	80.1	83.5	49.0
PAST∗[32]	ICCV19	78.4	-	-	54.6	72.4	-	-	54.3
SSG∗[5]	ICCV19	80.0	90.0	92.4	58.3	73.0	80.6	83.2	53.4
Ours (NPLP)	-	**89.4**	**95.4**	**96.8**	**74.3**	**78.5**	**87.0**	**89.7**	**61.3**

Table 4: Performance of NPLP on MSMT17.

Methods	Market1501 → MSMT17		
	Rank-1	Rank-10	mAP
PTGAN[29]	10.2	24.4	2.9
ECN[39]	25.3	42.1	8.5
SSG[5]	31.6	49.6	13.2
Ours (NPLP)	**52.8**	**69.5**	**23.3**

Methods	DukeMTMC-reID → MSMT17		
	Rank-1	Rank-10	mAP
PTGAN[29]	11.8	27.4	3.3
ECN[39]	30.2	46.8	10.2
SSG[5]	32.2	51.2	13.3
Ours (NPLP)	**52.6**	**69.6**	**23.3**

4.5 Comparison with State-of-the-art Methods

We compared our approach with state-of-the-art unsupervised person re-ID methods on Market1501, DukeMTMC-reID as well as MSMT17 in Table 3 and Table 4 respectively. It is clear that the performance of our method far beyond state-of-the-art methods to our best knowledge. We achieve **89.4% rank-1 accuracy and 74.3% mAP** on Market1501, which exceed the pseudo label estimation based methods[5, 15, 22, 32], by 9.4%-20% and 16%-36% respectively. Our method also has huge advantages on the DukeMTMC-reID dataset, surpassing the best published-method SSG [5] by 5.5% and 7.9% in rank-1 accuracy and mAP respectively. Specifically, we outperform the best-published method [5] by 21% in rank-1 accuracy and 10% in mAP when testing on MSMT17. It is worth noting that [5] utilizes the multi-granularity features independently, thus fail to explore the abundant features and robust correlations between samples. On the other hand, none of those pseudo label estimation based methods take pseudo label noise into account. Our method reduces pseudo label noise by training $\mathcal{M}$ iteratively and achieves state-of-the-art performance to the best of our knowledge.

5 CONCLUSION

In this paper, we analyze two types of pseudo label noise and their causes in pseudo label based unsupervised person re-ID. Then, we propose a Noisy Pseudo Label Processing(NPLP) framework to reduce noisy pseudo labels and train the re-ID model with noisy pseudo labels iteratively. We first adopt the Noise Reducing Strategy to construct a robust feature for pseudo label estimation. Then, we obtain pseudo labels with less noise for the re-ID model to learn. Meanwhile, we group those similar pairs to suppress the fragmented noise and train the re-ID model with pseudo labels that contain mixed noise at the Eager 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. Extensive experiments on three benchmark show that our noisy pseudo label processing framework significantly outperforms state-of-the-art unsupervised re-ID models by clear margins.

REFERENCES

[1] Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A Closer Look at Memorization in Deep Networks. In *ICML*.

[2] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*.

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*.

[4] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2018).

[5] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*.

[6] Aritra Ghosh, Himanshu Kumar, and P S Sastry. 2017. Robust Loss Functions under Label Noise for Deep Neural Networks.. In *AAAI*.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[9] Houjing Huang, Wenjie Yang, Xiaotang Chen, Xin Zhao, Kaiqi Huang, Jinbin Lin, Guan Huang, and Dalong Du. 2018. EANet: Enhancing Alignment for Cross-Domain Person Re-identification. *CoRR* (2018).

[10] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. 2019. O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks. In *ICCV*.

[11] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, Francis R. Bach and David M. Blei (Eds.).

[12] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. 2019. NLNL: Negative Learning for Noisy Labels. In *ICCV*.

[13] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. 2018. CleanNet: Transfer Learning for Scalable Image Classifier Training With Label Noise. In *CVPR*.

[14] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

[15] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. 2019. A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification. In *AAAI*.

[16] Zimo Liu, Dong Wang, and Huchuan Lu. 2017. Stepwise Metric Promotion for Unsupervised Video Person Re-identification. In *ICCV*.

[17] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. 2018. Unsupervised Cross-dataset Person Re-identification by Transfer Learning of Spatial-Temporal Patterns. In *CVPR*.

[18] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, Johannes Fürnkranz and Thorsten Joachims (Eds.).

[19] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *CVPR*.

[20] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. 2016. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*.

[21] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*.

[22] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. 2018. Unsupervised domain adaptive re-identification: Theory and practice. *arXiv preprint arXiv:1807.11334* (2018).

[23] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *ECCV*.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*.

[25] Arash Vahdat. 2017. Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks. In *NIPS*.

[26] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *ACM MM*.

[27] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 2018. Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification. In *CVPR*.

[28] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *ECCV*.

[29] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*.

[30] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. 2019. Patch-Based Discriminative Feature Learning for Unsupervised Person Re-Identification. In *CVPR*.

[31] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. 2019. Robust Person Re-Identification by Modelling Feature Uncertainty. (2019).

[32] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. 2019. Self-Training With Progressive Augmentation for Unsupervised Cross-Domain Person Re-Identification. In *(ICCV)*.

[33] Zhilu Zhang and Mert R Sabuncu. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *NIPS*.

[34] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. 2019. Pyramidal Person Re-IDentification via Multi-Loss Dynamic Training. In *CVPR*.

[35] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *ICCV*.

[36] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*.

[37] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2017. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896* (2017).

[38] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. 2018. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*.

[39] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*.

[40] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. 2018. Camera style adaptation for person re-identification. In *CVPR*.