Delving into Probabilistic Uncertainty for Unsupervised Domain Adaptive Person Re-identification

Jian Han, Ya-Li Li*, Shengjin Wang
Beijing National Research Center for Information Science and Technology (BNRist)
Department of Electronic Engineering, Tsinghua University
hanj19@mails.tsinghua.edu.cn, {liyali13, wgsgj}@tsinghua.edu.cn

Abstract
Clustering-based unsupervised domain adaptive (UDA) person re-identification (ReID) reduces exhaustive annotations. However, owing to unsatisfactory feature embedding and imperfect clustering, pseudo labels for target domain data inherently contain an unknown proportion of wrong ones, which would mislead feature learning. In this paper, we propose an approach named probabilistic uncertainty guided progressive label refinery (P^2LR) for domain adaptive person re-identification. First, we propose to model the labeling uncertainty with the probabilistic distance along with ideal single-peak distributions. A quantitative criterion is established to measure the uncertainty of pseudo labels and facilitate the network training. Second, we explore a progressive strategy for refining pseudo labels. With the uncertainty-guided alternative optimization, we balance between the exploration of target domain data and the negative effects of noisy labeling. On top of a strong baseline, we obtain significant improvements and achieve the state-of-the-art performance on four UDA ReID benchmarks. Specifically, our method outperforms the baseline by 6.5% mAP on the Duke2Market task, while surpassing the state-of-the-art method by 2.5% mAP on the Market2MSMT task. Code is available at: https://github.com/JeyesHan/P2LR.

Introduction
Person re-identification (ReID) aims to retrieve all images of the target person from non-overlapping camera views. It has broad applications in smart retail, searching missing children, and other person related scenarios. Although existing deep learning methods (Zhang, Zhang, and Liu 2021; Isobe et al. 2020; Luo et al. 2019; Guo et al. 2019; Hermans, Beyer, and Leibe 2017) have achieved remarkable performance, these methods rely heavily on manual annotations. Furthermore, they fail to generalize on other datasets when domain gap between the source and target domain exists. To address this critical issue, UDA person ReID where labeled source domain images and unlabeled target images are presented has attracted great attention in recent years.

UDA person ReID methods include domain transfer based methods (Huang et al. 2020b; Ge et al. 2020; Deng et al. 2018; Wei et al. 2018; Zhong et al. 2018), ranking based methods (Zhong et al. 2019; Wang et al. 2020; Yu et al. 2019), and clustering based methods (Fan et al. 2018; Lin et al. 2019; Fu et al. 2019; Ge, Chen, and Li 2020; Zhai et al. 2020b). In general, clustering based methods can obtain superior performance. Typically, clustering based methods consist of three stages, i.e., i) pre-training with labeled source domain images, ii) generating pseudo labels for target domain images by clustering, and iii) fine-tuning with target domain images. The first stage is conducted only once and the latter two stages are repeated several times for mutual promotion. However, owing to unsatisfactory feature embedding and imperfect clustering quality, we are supposed to tackle an unknown proportion of wrong pseudo labels. On one hand, such noisy labels inherent with cross-domain pseudo labeling would mislead the network optimization in the fine-tuning stage. On the other hand, the negative effects caused by wrong pseudo labels will propagate and amplify as the training procedure proceeds. Therefore, it is crucial to identify wrong pseudo labels and alleviate the uncertainty of pseudo labeling to promote the UDA person ReID.

This paper addresses two key issues for clustering-based methods: 1) how to identify the wrong pseudo labels, 2) how to reduce the negative effects of wrongly-labeled samples in optimization. For the first issue, we observe that the probability distribution among identities of wrong-labeled sam-
samples is significantly different from that of samples with correct pseudo labels. Samples with correct pseudo labels are explicit, reaching high in the corresponding true identity and remain close to zero in other identities. In contrast, samples with wrong pseudo labels are ambiguous, with several peaks in several other identities. This obvious difference inspires us to model the probabilistic uncertainty by measuring the inconsistency between the predicted and ideal distributions of pseudo labels (Figure 1). For the second issue, we propose a probabilistic uncertainty guided progressive label refinery framework and solve it by Alternative Convex Search (Gorski, Pfeuffer, and Klamroth 2007). We alternatively include samples with high credibility and minimize the uncertainty of selected samples to boost the performance of domain adaptive person ReID.

Our main contributions can be summarized as follows:

1. **We propose to measure the probabilistic uncertainty of pseudo labels for UDA person ReID.** A quantitative criterion that estimates the probabilistic distance of the predicted distribution of samples against ideal ones, is established to serve as the standard for eliminating wrongly-labeled samples.

2. **We propose a framework named Probabilistic uncertainty guided Progressive Label Refinery (P^2LR) to determine and purify wrong labels of target domain samples for UDA person ReID.**

3. **We conduct extensive experiments and achieve the state-of-the-art performance on four benchmark datasets.** P^2LR outperforms the baseline by 6.5% mAP on the Duke2Market task and surpasses the state-of-the-art method by 2.5% mAP on the Market2MSMT task.

### Related Work

**UDA person ReID.** To save exhaustive annotations, unsupervised domain adaptation (UDA) (Wei et al. 2021; Isobe et al. 2021; Cui et al. 2020; Gong et al. 2019) makes models trained on source domain adapt to target domain. Existing UDA ReID methods can be divided into three categories, *i.e.*, domain transfer based methods, ranking based methods, and clustering based methods.

**Domain transfer based methods.** (Huang et al. 2020b; Ge et al. 2020) leverage style-transfer techniques to transfer source images into target domain. Then the transferred images with inherited labels are utilized to fine-tune the model pre-trained on source domain. SPGAN (Deng et al. 2018) and PTGAN (Wei et al. 2018) use GANs to transform source-domain images to match the image styles of the target domain. HHL (Zhong et al. 2018) proposes to learn camera-invariant embeddings via transforming images according to target camera styles. However, the image generation quality is still unsatisfactory and the information of target domain is not fully explored because target domain images are only used for providing style supervision signals.

**Ranking based methods build positive and negative sample pairs and use the contrastive loss to optimize neural networks with soft labels on the target domain.** ECN (Zhong et al. 2019) builds a memory bank to store seen features. Exemplar, camera, and neighborhood invariance are modeled by pair learning with features in the memory bank. Wang et al. (Wang et al. 2020) use memory bank to mine hard negative instances across batches. MAR (Yu et al. 2019) exploits a set of reference persons to generate multiple soft labels. However, the constructed image pairs or feature pairs lacks global guidance over all identities. Additionally, the reference images or features might be outliers in some cases which mislead the feature learning.

**Clustering based methods obtain superior performance to domain transfer based and ranking based methods to date.** Clustering methods generate pseudo labels through clustering and then fine-tune models with generated pseudo labels. Fan et al. (Fan et al. 2018) propose to alternatively assign labels for unlabeled training samples and optimize the network with the generated targets. BUC (Lin et al. 2019) proposes a bottom-up clustering framework to gradually group similar clusters. SSG (Fu et al. 2019) employs both global body and local body part features for clustering and evaluation. DAAM (Huang et al. 2020a) introduces domain alignment constraints and an attention module. MMT (Ge, Chen, and Li 2020) introduces two sibling mutual mean teachers. MEB-Net (Zhai et al. 2020b) establishes three networks to perform mutual mean learning. However, these methods ignore noisy/wrong labels generated by clustering methods, which hinders the advancement of these approaches.

**Pseudo label evaluation.** Wrong pseudo label samples harm learning robust feature embedding for neural networks. To quantify and identify the correctness of pseudo labels, pseudo label evaluation becomes significantly crucial. Kendall et al. (Kendall and Gal 2017) and Chang et al. (Chang et al. 2020) establish an end-to-end framework to measure the observation noise and alleviate the negative influence for better network optimization. Zheng et al. (Zheng and Yang 2021) propose to estimate the correctness of predicted pseudo labels in semantic segmentation. As for clustering-based UDA person ReID, EUG (Wu et al. 2019), UNRN (Zheng et al. 2021a), and GLT (Zheng et al. 2021b) are uncertainty based methods. EUG employs the L2 distance between samples and clustering centroids in feature space to determine the reliability of samples. UNRN measures the output consistency between mean teacher and student models as uncertainty, while we model the distribution of single-model clusters and measure the probabilistic distance to an ideal distribution as the uncertainty. GLT is a group-aware label transfer framework to explicitly correct noisy labels while we pick reliable pseudo labels to train the model progressively, which further correct noisy labels implicitly.

### Methodology

**Notation.** The goal of UDA person ReID is to adapt the trained model from a source domain \(\mathbb{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}\) to an unsupervised target domain \(\mathbb{D} = \{x_i\}_{i=1}^{N}\), where \(N_s\) and \(N\) denote the number of samples in the source and target domains, respectively. \(x_i^s\) and \(y_i^s\) denote the sample and its attached label in the source domain with supervisory information. \(x_i\) denotes the sample in target domain without supervision. The pseudo label generated by clustering for sample \(x_i\) in target is denoted as \(\tilde{y}_i\).
Figure 2 shows an overview of our proposed P²LR framework for UDA person ReID. We first construct a clustering baseline with mutual mean teachers (Ge, Chen, and Li 2020). On top of the baseline, we introduce the probabilistic uncertainty guided progressive label refinery to evaluate the noisy level of pseudo labels and reduce the negative influence of noisy samples. In the following parts, we first introduce the clustering baseline, and then elaborate on the label uncertainty modeling based on probabilistic uncertainty and the proposed P²LR framework.

Recap Clustering Baseline

We construct a clustering-based pipeline based on MMT (Ge, Chen, and Li 2020) for UDA person ReID. Following the general pipeline of clustering-based methods, three stages are consisted, such as: source-domain model pretraining, clustering, and target-domain fine-tuning. In the source pre-training stage, two sibling networks are established with the same architecture. They are initialized with different random seeds, trained with the same data yet experienced different data augmentations to reduce the dependence. The networks are optimized by the identity loss $L_{id}$ and triplet loss $L_{tri}$. In the clustering stage, pseudo labels for target-domain data are generated by clustering. In the target fine-tuning stage, two mean teachers $\bar{M}_1$ and $\bar{M}_2$ are established with the exponential moving average of student models over iterations. The prediction of one teacher model is taken as soft labels to mutually supervise the training of the other student. Apart from the identification and triplet loss with hard pseudo labels provided by clustering, each model is also guided by the Kullback-Leibler (KL) divergence loss $L_{KL}$ and soft triplet loss $L_{tri}$ with the soft labels.

The whole loss function in the fine-tuning stage is as:

$$L = L_{id} + \lambda_{tri}L_{tri} + \lambda_{KL}L_{KL} + \lambda_{tri}L_{tri}$$

(1)

where $\lambda_{tri}$, $\lambda_{KL}$, $\lambda_{tri}$ indicate corresponding loss weights.

Probabilistic Uncertainty Modeling

For clustering-based UDA person ReID, the pseudo labels are noisy, which would mislead the network training in the target fine-tuning stage and hurt the performance in target domain. To reduce the effects of noise, uncertainty estimation is a natural way to eliminate the unreliable pseudo labels. To establish the uncertainty criterion, we present the output probabilistic distribution of samples with correct and wrong labels for an investigative study. As shown in Figure 3a, the sample with wrong pseudo labels is associated with high probabilistic uncertainty, which has several peaks and generally reaches high in different identities. In contrast, the sample with correct pseudo labels is associated with low probabilistic uncertainty (Figure 3b). The assigned probability peaks high in the correct identity and remains nearly zero in other identities. This is consistent with the intuitive idea that the samples with wrong pseudo labels are ambiguous for model prediction, which is significantly distinct to correct pseudo labels.

Inspired by this observation, we leverage the distribution difference as the probabilistic uncertainty to softly evaluate the noisy level of samples. Each unlabeled sample $x_i$ in the target domain is assigned with a pseudo label $\tilde{y}_i$ by clustering. Based on the clustering at $t$-th step, we build an external classifier $\phi_t(\cdot|w_t^{cls})$, where $w_t^{cls} \in \mathbb{R}^{c \times d}$ is the parameterized weights of the external classifier. Note that the weights are generated dynamically and don’t need separate training. Here we employ $c$ cluster centroids of $d$ dimensions as classifier weights. For a feature $f_i \in \mathbb{R}^d$ of sample $(x_i, \tilde{y}_i)$ extracted by the mean teacher, we obtain classified probability distributions among identities with this classifier following Eq.2. Here $\alpha$ is the temperature parameter.

$$P(x_i, \tilde{y}_i) = \phi_t(f_i) = \text{Softmax} \left( \frac{w_t^{cls}}{||w_t^{cls}||} \cdot \frac{f_i}{||f_i||}, \alpha \right)$$

(2)
In source domain, we obtain a single impulse distribution for samples after symmetrically arranging the probability from large to small while making $y_i$ as the center. This phenomenon motivates us to model the generalized ideal distributions for samples in the target domain. We find that the smoothed $\delta$ distribution associative with large temperature $\alpha$ is highly stable and insensitive to specific datasets. When $\alpha$ decreases (e.g., controls the variety of distributions), another distribution might be better for some cases but hard to generalize. Therefore, we draw the smoothed $\delta$ distribution (Eq. 3) as the ideal distribution $Q(x_i, \tilde{y}_i)$.

$$Q(x_i, \tilde{y}_i) = \delta_{\text{smooth}}(j - \tilde{y}_i) = \begin{cases} 
\epsilon, & \text{if } j = \tilde{y}_i \\
\frac{1 - \epsilon}{c - 1}, & \text{otherwise}
\end{cases} \quad (3)$$

where $j$ is the identity index and $\epsilon$ is the number of identities (i.e., the number of clusters). $\epsilon$ is a hyperparameter and is set to 0.99.

We estimate the probabilistic distance of the predicted probability across identities and the ideal distribution to measure the uncertainty of samples. Rather than measuring the feature inconsistency of teacher-student models (Zheng et al. 2021a), we define a criterion named probabilistic uncertainty, which measures the inconsistency between predicted distribution $P(x_i, \tilde{y}_i)$ and the ideal distribution $Q(x_i, \tilde{y}_i)$.

Moreover, we leverage the Kullback–Leibler (KL) divergence to measure the inconsistency and establish the probabilistic uncertainty as:

$$U(x_i, \tilde{y}_i) = D_{KL}(Q(x_i, \tilde{y}_i) \| P(x_i, \tilde{y}_i)) \quad (4)$$

Larger $U(x_i, \tilde{y}_i)$ indicates that the pseudo label generated by clustering is more likely to be wrong, which should be excluded in the target fine-tuning stages. As in Figure 3c, our probabilistic uncertainty is highly related to correctness.

### Uncertainty Guided Alternative Optimization

We formulate the uncertainty guided domain adaption mathematically. The model can be denoted as $M(\cdot | w)$, where $w$ is the associated parameters. Each model maps a sample $x_i$ to the prediction $M(x_i | w)$. An uncertainty score can be further obtained, as $u_i = U(x_i, \tilde{y}_i | w)$. $U$ is the label uncertainty estimation module. The uncertainty guided learning can be formulated as a joint optimization problem as:

$$\min_{w, v, \beta} \mathbb{E}(w, v; \beta) = \sum_{i=1}^{N} v_i u_i - \beta \sum_{i=1}^{N} v_i \quad (5)$$

where $u_i \in \{0, 1\}$ is the indicator for sample selection. $\beta$ is the age parameter to control the learning pace. That is, we need to reduce the overall uncertainty of select samples while selecting as many samples as possible for sufficient training. To optimize the above objective function, we adopt Alternative Convex Search (Gorski, Pfeuffer, and Klamroth 2007). In particular, $w$ and $v$ are alternatively optimized while fix the other. With the fixed $w^*$, the global optimum $v^*$ is solved as:

$$v^* = \arg \min_{v_i \in \{0, 1\}} \sum_{i=1}^{N} v_i u_i - \beta \sum_{i=1}^{N} v_i \quad (6)$$

When $v^*$ is fixed, the global optimum $w^*$ is solved as:

$$w^* = \arg \min_{w} \sum_{i=1}^{N} v_i^* u_i \quad (7)$$

The two optimization steps are iteratively conducted, while $\beta$ is gradually increased to add more harder samples.

**Uncertainty guided sample selection.** With the proposed uncertainty measurement, the solution for Eq.7 is provided by gradient descent based network training. With fixed $w^*$, finding the optimal $v$ is converted into a combinatorial optimization issue with the linear objective function. We can simply derive the global minimum by setting the partial
derivative of Eq.6 to \( v_t \) as zero. Considering \( v_t \) is either 0 or 1, we obtain the close-formed solution \( v^* \) as:

\[
v^*_i = \begin{cases} 
1, & \text{if } u_i \leq \beta \\
0, & \text{otherwise}
\end{cases}
\]  

(8)

Since clustering and target fine-tuning stages are iterated in an alternative way for multiple steps, we need to determine how many samples are included. A sequence \( \{N \cdot p_t\}, t = 0, 1, \cdots, T \) is predefined for this purpose, where \( N \cdot p_t \) is the number of selected examples at iteration time \( t \). Based on the predefined \( p_t \), the threshold \( \beta \) is dynamically updated according to Eq.9 to ensure exactly \( N \cdot p_t \) samples are assigned with non-zero \( v_i \).

\[
\beta = \text{Sort}\{u_1, u_2, \ldots, u_N\}(N \cdot p_t)
\]  

(9)

Note that as the training proceeds, the models become more reliable. Consequently, the measured uncertainty reduces drastically at the beginning, then converges slowly with larger \( t \) (Figure 4(left)). As the result, \( p_t \) should be enlarged as \( t \) increases. To be consistent with the trends of uncertainty reduction, we design \( p_t \) as a rescaled logarithm-exponential function:

\[
p_t = p(t) = p_0 + \frac{1}{h} \ln \left( 1 + \left(e^{h \cdot (1 - p_0)} - 1\right) \cdot \frac{t}{T} \right)
\]  

(10)

\( p_0 \) is the initial proportion of samples for fine-tuning in target domain. The setting of \( p_0 \) enables the model to learn from sufficient easy samples at early steps, to avoid falling into the negative loop of degrading. \( h \) is the hyper-parameter to control the increasing rate of samples (Figure 4(right)). Because the rate of adding examples is negatively correlated to the size of the current training set, the number of newly added examples is reduced as the training proceeds. More importantly, the uncertainty decreases dramatically when \( t \) is small, we enlarge the set of samples in a fast way. As \( t \) increases, the uncertainty converges and the set of samples for fine-tuning increases slowly for stable training. This progressive label refinery would help the model learn sufficiently from the newly added samples in target domain.

Experiments

Implementation Details

We implement our model based on MMT (Ge, Chen, and Li 2020) and train it on four Tesla XP GPUs. ADAM optimizer is adopted to optimize models with the weight decay of \( 5 \times 10^{-4} \). We employ ImageNet (Deng et al. 2009) pre-trained ResNet50 (He et al. 2016) as the backbone. We obey normal \( P - K \) sampling in person ReID, where \( P = 4 \) and \( K = 16 \) in each mini-batch. Then we perform data augmentation of random cropping, flipping, and erasing. Note that random erasing is not utilized in the source pre-training stage. All person images are resized to \( 256 \times 128 \). We use the clustering algorithm of k-means where the number of clusters (c) is set as 500, 700, and 1500 for Market, Duke, and MSMT datasets, respectively. The temperature parameter \( \alpha \) in Eq.2 is set to 20. The parameter \( p_0 \) and \( h \) in Eq.10 are set to 0.3 and 1.5, respectively. Overall alternative optimization steps \( T \) are set to 100. In source pre-training stage, the initial learning rate is set to \( 3.5 \times 10^{-4} \) and is decreased by 1/10 on the 40th and 70th epoch in the total 80 epochs. In target fine-tuning stage, the learning rate is fixed to \( 3.5 \times 10^{-4} \).

Datasets and Protocols

We evaluate our method on three main-stream person ReID datasets, i.e., Market-1501 (Market) (Zheng et al. 2015), DukeMTMC-reID (Duke) (Ristani et al. 2016), and MSMT17 (MSMT) (Wei et al. 2018). The Market-1501 dataset contains 32,668 annotated images of 1,501 identities shot from 6 cameras, where 12,936 images of 751 identities are used for training and 19,732 images of 750 identities for testing. The DukeMTMC-reID dataset consists of 36,411 images collected from 8 cameras, where 702 identities are used for training and 702 identities for testing. The MSMT17 is the most challenging and largest person ReID dataset consisting of 126,441 images of 4,101 identities, where 1,041 identities and 3,060 identities are used for training and testing, respectively. We adopt mean average precision (mAP) and CMC Rank-1/5/10 (R1/R5/R10) accuracy without re-ranking (Zhong et al. 2017) for evaluation.

Comparison with State-of-the-Arts

We compare our method against the state-of-the-art (SOTA) methods on four UDA ReID settings and present the results in Table 1. Among existing methods for UDA person ReID, DAAM (Huang et al. 2020a) introduces domain alignment constraints and an attention module. SSG (Fu et al. 2019), MMT (Ge, Chen, and Li 2020), MEB-Net (Zhai et al. 2020b), and UNRN (Zheng et al. 2021a) are all clustering-based methods. SSG (Fu et al. 2019) employs both global body and local body part features for clustering and evaluation. We construct the baseline based on MMT (Ge, Chen, and Li 2020) which introduces mutual mean teacher for UDA person ReID. Compared to the baseline MMT (Ge, Chen, and Li 2020), our proposed \( P^2 LR \) significantly improves the UDA ReID accuracy, with 9.8%, 5.7%, 6.1%, and 6.6% mAP improvements on four UDA ReID settings. Compared to MEB-Net (Zhai et al. 2020b) which establishes three networks to perform mutual mean learning, we increase the mAP by 5.0%, 4.7% with a simpler architecture design. Notably, UNRN and GLT leverage source data during target fine-tuning stage and build an external support memory to mine hard pairs. Our \( P^2 LR \) still achieves 3.7% and 3.7% mAP gains to UNRN, 2.5% and 2.2% mAP gains to GLT on the MSMT dataset. In general, our method \( P^2 LR \)


<table>
<thead>
<tr>
<th>Methods</th>
<th>DukeMTMC $\rightarrow$ Market1501</th>
<th>Market1501 $\rightarrow$ DukeMTMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>R1</td>
<td>R5</td>
</tr>
<tr>
<td>ATNet (Liu et al. 2019)(CVPR’19)</td>
<td>25.6</td>
<td>55.7</td>
</tr>
<tr>
<td>SPGAN+LMP (Deng et al. 2018)(CVPR’18)</td>
<td>26.7</td>
<td>57.7</td>
</tr>
<tr>
<td>BUC (Lin et al. 2019) (AAAI’19)</td>
<td>38.3</td>
<td>66.2</td>
</tr>
<tr>
<td>ECN (Zhong et al. 2019) (CVPR’19)</td>
<td>43.0</td>
<td>75.1</td>
</tr>
<tr>
<td>PDA-PAST (Li et al. 2019) (ICCV’19)</td>
<td>47.6</td>
<td>75.2</td>
</tr>
<tr>
<td>PCB-PAST (Zhang et al. 2019) (ICCV’19)</td>
<td>54.6</td>
<td>78.4</td>
</tr>
<tr>
<td>SSG (Fu et al. 2019) (ICCV’19)</td>
<td>58.3</td>
<td>80.0</td>
</tr>
<tr>
<td>ACT (Yang et al. 2020) (AAAI’20)</td>
<td>60.6</td>
<td>80.5</td>
</tr>
<tr>
<td>MQLP (Wang and Zhang 2020) (CVPR’20)</td>
<td>60.4</td>
<td>84.4</td>
</tr>
<tr>
<td>DAAM (Huang et al. 2020a) (AAAI’20)</td>
<td>67.8</td>
<td>86.4</td>
</tr>
<tr>
<td>AD-Cluster (Zhai et al. 2020a) (CVPR’20)</td>
<td>68.3</td>
<td>86.7</td>
</tr>
<tr>
<td>MMT (Ge, Chen, and Li 2020) (ICLR’20)</td>
<td>71.2</td>
<td>87.7</td>
</tr>
<tr>
<td>NRMT (Zhao et al. 2020)(ECCV’20)</td>
<td>71.7</td>
<td>87.8</td>
</tr>
<tr>
<td>B-SNR+GDS-H (Jin et al. 2020)(ECCV’20)</td>
<td>72.5</td>
<td>89.3</td>
</tr>
<tr>
<td>MEB-Net (Zhai et al. 2020b)(ECCV’20)</td>
<td>76.0</td>
<td>89.9</td>
</tr>
<tr>
<td>UNRN (Zheng et al. 2021a) (AAAI’21)</td>
<td>78.1</td>
<td>91.9</td>
</tr>
<tr>
<td>GLT (Zheng et al. 2021b)(CVPR’21)</td>
<td>79.5</td>
<td>92.2</td>
</tr>
<tr>
<td><strong>P</strong>$^2$LR (Ours)</td>
<td><strong>81.0</strong></td>
<td><strong>92.6</strong></td>
</tr>
</tbody>
</table>

Table 1: Performance (%) comparison with the state-of-the-art methods for UDA person ReID on the datasets of DukeMTMC-reID, Market-1501 and MSMT17. Results of the best are marked by bold text.

achieves the state-of-the-art performance on all datasets, which verifies the effectiveness of **P**$^2$LR.

Ablation Studies

In this section, we carry out extensive ablation studies to validate the effectiveness of each component in **P**$^2$LR.

Effectiveness of progressive label refinery. To alleviate the negative effects of noisy/wrong pseudo labels, we exploit the characteristics of probability distribution among identities to evaluate probabilistic uncertainty of pseudo labels and refine the pseudo labels by alternative optimization. We validate the effectiveness of the proposed progressive label refinery module by comparing the performance of Baseline and Baseline+**P**$^2$LR. Without **P**$^2$LR, the achieved mAP and Rank-1 accuracy are 74.5% and 90.3%, respectively. With **P**$^2$LR, the achieved mAP is 81.0% and Rank-1 accuracy is 92.6%, respectively. The mAP is significantly improved by 6.5% with the sole **P**$^2$LR on the Duke2Market task. The results show that the proposed method is effective to alleviate the negative effects caused by noisy/wrong pseudo labels and significantly improves UDA person ReID performance.

Different designs of label correctness estimation. A straightforward way to evaluate the uncertainty of pseudo labels is referring to the $l_2$ distance to the nearest centroid in feature space. As shown in Table 2, the mAP and Rank-1 accuracy is improved by 3.2% and 1.8% with the probabilistic uncertainty criterion. That is because the probability distribution based uncertainty evaluation considers more global references than $l_2$ distance in local feature space. Besides, we find introducing an external classifier with cluster centred for uncertainty evaluation outperforms internal classifiers in $M(l \cdot w)$. It can be attributed to the fact that internal classifiers are more easily misled by noisy labeling during training. External classifiers whose weights are generated by clustering could correct the misleading to some extent.

We compare our label evaluation method to prediction consistency based method UNRN. Our method surpasses UNRN with large margins. The uncertainty (noise level) of pseudo labels should be measured from the intrinsic property of its own probabilistic distribution, especially for UDA ReID. Since noise already exists in teacher models, the consistency-based uncertainty as in UNRN might overfit biased teacher models.

Progressive label refinery vs Sample Reweighting. We compare the label refinery with reweighting method proposed in UNRN (Zheng et al. 2021a). In general, our method
Figure 5: Visualization of refined pseudo label samples. We present samples with high certainty and uncertain samples of three clusters measured in the first refinery step. We can observe that most certain samples are from the same identity while uncertain samples are hard samples of the identity or from other identities.

Table 2: Ablation studies on the effectiveness of components in the proposed method on the Duke2Market benchmark.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Duke→Market</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP</td>
</tr>
<tr>
<td>Model pretraining</td>
<td>29.6</td>
</tr>
<tr>
<td>Base.</td>
<td>58.0</td>
</tr>
<tr>
<td>Base.+ML</td>
<td>68.5</td>
</tr>
<tr>
<td>Base.+ML+MT (Baseline)</td>
<td>74.5</td>
</tr>
<tr>
<td>Baseline + $l_2$ distance</td>
<td>77.8</td>
</tr>
<tr>
<td>Baseline + Internal classifier</td>
<td>80.2</td>
</tr>
<tr>
<td>Baseline + Consistency</td>
<td>79.2</td>
</tr>
<tr>
<td>Baseline + Reweighting</td>
<td>77.7</td>
</tr>
<tr>
<td>Baseline + $P^2$LR($\epsilon$=0.95)</td>
<td>74.8</td>
</tr>
<tr>
<td>Baseline + $P^2$LR($\epsilon$=0.97)</td>
<td>77.1</td>
</tr>
<tr>
<td>Baseline + $P^2$LR($\epsilon$=1.00)</td>
<td>80.8</td>
</tr>
<tr>
<td>Baseline + $P^2$LR (Ours)</td>
<td>81.0</td>
</tr>
</tbody>
</table>

Filters out wrong labels with the guidance of uncertainty, while the reweighting-based method reduces the negative influence in loss functions. The comparative results are presented in Table 2. The proposed progressive label refinery significantly outperforms reweighting method. It indicates that removing wrong labeled samples in training is more effective for reducing their negative influence.

Parameter analysis. We experiment with different ideal distribution $Q$ by setting different $\epsilon$. We present the mAP/R1/R5/R10/R10 in Table 2 and mAP curves with the number of refinery steps in Figure 6a, respectively. It can be seen $\epsilon$ has important impacts while the mAP curve of $\epsilon$=0.99 grows fastest and highest. Therefore, we set $\epsilon$=0.99 for all other experiments. We then explore the influence of temperature parameter $\alpha$ in the uncertainty measurement while setting $p_0 = 0.2$ in Figure 6b. The achieved mAP and Rank-1 accuracy maintain high when $\alpha$ is set to 20 ~ 50. When $\alpha = 20$, the achieved mAP is 80.4% and Rank-1 accuracy is 91.4%. When $\alpha = 50$, the achieved mAP is 80.2% and Rank-1 accuracy is 92.0%. Afterwards, we present the influence of $p_0$ in progressive label refinery Figure 6c setting $\alpha = 20$. The mAP and Rank-1 accuracy reach peak (81.0% and 92.6%) when $p$ is around 0.3. It indicates we preserve about 30% of samples and eliminate 70% ones in the early refinery steps. We progressively exploit more uncertain samples in later refinery steps for strong performance.

Visualization. We present the visualization results to validate the effectiveness of probabilistic uncertainty measurement for selecting samples. The selected certain samples and un-selected uncertain samples of three clusters determined by the proposed criterion in the first refinery step are presented in Figure 5. It can be observed that most selected samples are from the same identity while un-selected samples are hard samples of the identity or from other identities. By combining the selected certain samples with high confidence into target domain fine-tuning, the proposed $P^2$LR can achieve superior UDA person ReID performance.

Conclusion

In this paper, we aim to find wrong pseudo labels and reduce their negative effects for clustering-based UDA person ReID. We observe that the probability distributions of samples with wrong pseudo labels are ambiguous with multiple peaks, which is different from that of correctly-labeled samples. This inspires us to propose a probabilistic uncertainty guided progressive label refinery ($P^2$LR) framework for UDA person ReID. We solve it by alternative optimization, which balances between the target domain exploitation and noise label overfitting. The proposed method brings significant improvements over a strong baseline and achieves the state-of-the-art performance on four benchmarks.
Acknowledgments

This work was supported by the Cross-Media Intelligent Technology Project of Beijing National Research Center for Information Science and Technology (BNRist) under Grant No.BNR2019TD01022, National Natural Science Foundation of China under Grant No.61771288, and the research fund under Grant No. 2019GGG0001 from the Institute of GuoQiang, Tsinghua University.

References


