Reliability Exploration with Self-ensemble Learning for Domain Adaptive Person Re-Identification

Zongyi Li, Yuxuan Shi*, Hefei Ling, Jiazhong Chen, Qian Wang, Fengfan Zhou

Department of Computer Science and Technology, Huazhong University of Science and Technology
{zongyili, shiyx, lhefei, jzchen, yqwq1996, fengfanzhou}@hust.edu.cn

Abstract

Person re-identification (Re-ID) based on unsupervised domain adaptation (UDA) aims to transfer the pre-trained model from one labeled source domain to an unlabeled target domain. Existing methods tackle this problem by using clustering methods to generate pseudo labels. However, pseudo labels produced by these techniques may be unstable and noisy, substantially deteriorating models' performance. In this paper, we propose a Reliability Exploration with Self-ensemble Learning (RESL) framework for domain adaptive person Re-ID. First, to increase the feature diversity, multiple branches are presented to extract features from different data augmentations. Taking the temporally average model as a mean teacher model, online label refining is conducted by using its dynamic ensemble predictions from different branches as soft labels. Second, to combat the adverse effects of unreliable samples in clusters, sample reliability is estimated by evaluating the consistency of different clusters' results, followed by selecting reliable instances for training and re-weighting sample contribution within Re-ID losses. A contrastive loss is also utilized with cluster-level memory features which are updated by the mean feature. The experiments demonstrate that our method can significantly surpass the state-of-the-art performance on the unsupervised domain adaptive person Re-ID.

Introduction

Person re-identification (Re-ID) aims to retrieve a given person from different cameras. With the development of deep learning and increasing computing power, person Re-ID has achieved great success (Zheng et al. 2019; Lin et al. 2019b; Shi et al. 2020a,b,c). However, most existing methods based on supervised learning require a large amount of labeled data which is time-consuming. And if a pre-trained model is transferred from one scenario to another, the model performance will drop drastically due to different data distribution. Therefore, the unsupervised domain adaption (UDA) for person Re-ID needs to be introduced for transferring a trained model from a labeled source domain to another unlabeled dataset.

Recently, most existing methods focusing on unsupervised domain adaption for person Re-ID can be broadly grouped into two categories. The methods (Bak, Carr, and Lalonde 2018; Deng et al. 2018; Wei et al. 2018; Zhong et al. 2018a; Li et al. 2019) in the first category align the image distributions between source and target datasets or apply Generative Adversarial Network to transfer the style of image from the source domain to the target domain, remaining the identity label of the image for training. However, these methods achieve unsatisfying performance because they do not make full use of the information of the target domain data. Most existing methods(Ding et al. 2020; Ge, Chen, and Li 2020; Zhai et al. 2020b; Zheng et al. 2021a) are based on the second category, named clustering-based methods. This kind of method utilizes the pre-trained model from the source domain to generate pseudo labels on the target domain by extracting features through clustering algorithms. The pseudo label generation step and the training step are performed alternatively until the model converges. Although these methods can achieve acceptable performance with pseudo labels, they also face the noisy problem due to the domain gap.

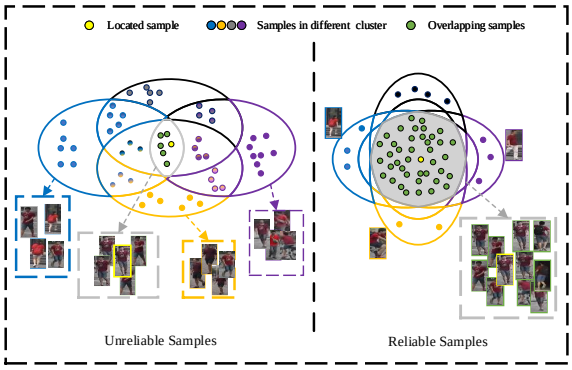

Figure 1: Illustration of the proposed reliability estimation method, which utilizes the overlapping sample in different cluster results. Different color circles represent different clusters produced by different branch features. A reliable instance is supposed to have more overlapping cluster neighbors, and those of a unreliable sample are scatted in different clusters.

To handle noisy labels, some researchers use mutual teaching (Zhang et al. 2018; Shi et al. 2021) to train paired networks and help to correct each other. MMT(Ge, Chen,

*Corresponding author

and Li 2020) trains two paired networks and corrects their pseudo labels by using their moving average networks. However, this training method would result in the two models fitting to each other. Moreover, MEB-Net(Zhai et al. 2020b) utilizes multiple networks with different architectures to enhance the feature diversity and try to correct noisy labels via a brainstorming training strategy. But this method requires training multiple models iteratively, which is hard for training. In this paper, a reliability exploration method is proposed with a self-ensemble learning framework, where online pseudo labels refining and offline reliable sample selection are utilized to handle the noise problem. Furthermore, our model consists of several branches after the stem CNN. To equip different branches with diverse knowledge and specific features, CycleGAN(Zhu et al. 2017) is adopted to generate images with different camera styles. Then, the predictions of each branch are integrated via a dynamic router. Meanwhile, the temporally averaged model is used as the teacher model, and its integrated prediction is employed as soft labels to supervise the learning of different branches, which can be regarded as a self-ensemble learning way. Compared with other mutual teaching methods, our model can avoid asynchronous model updating while achieving superior performance.

In addition, we propose a method to estimate the pseudo label reliability by calculating the consistency of different cluster results, which is shown in Fig.1. Specifically, different branch features in the running model and ensemble features in the temporally average model are utilized to cluster images in the target domain. Due to the diversity in features, the result of each cluster varies. The reliability of each sample is evaluated by calculating the intersection-over-union of different clusters in which the sample is located. By selecting instances with reliable labels and incorporating the reliability into the Re-ID loss for model training, we can relieve the negative influence of noisy pseudo labels.

Although self-ensemble learning and reliable sample selection can significantly reduce the influence of noisy pseudo labels, the reliability-aware classification loss and contrastive loss are also utilized to enhance the feature discrimination. Instead of saving the current feature in the memory bank like previous methods, the ensemble feature is applied to update the cluster-level memory bank in the temporally average model. The cluster centroids are used to guide different branch features through a multi-branch contrastive loss. Our contributions can be summarized as follows:

1. A multi-branch architecture with self-ensemble learning is proposed, which adopts self-ensemble prediction to increase the quality of mean teacher-based soft labels.

2. A sample reliability estimation strategy is presented by evaluating the consistency of sample clusters. By selecting reliable samples for training and incorporating reliability into the Re-ID loss, our model can mitigate the negative influence of noisy pseudo labels

3. Cluster-level memory bank is constructed with features in the temporally average average model, and the multi-branch contrastive learning is used to supervise different branches' features, which dramatically strengthens the discrimination of features.

Experiments show that our method achieves a tradeable effect and surpasses most state-of-the-art methods by large margins on multiple benchmarks of unsupervised domain adaptive Re-ID.

Related work

Fully unsupervised Re-ID

Fully unsupervised Re-ID (Lin et al. 2019a; Wang and Zhang 2020; Zeng et al. 2020; Ge et al. 2020) aims to train Re-ID model without any annotations. These methods mainly adopt cluster algorithm (Ester et al. 1996a) to generate pseudo label for model training. A bottom-up clustering framework (BUC) (Lin et al. 2019a) treats each independent sample as a cluster and then hierarchically groups similar clusters into one cluster to generate pseudo labels. HCT (Zeng et al. 2020) uses the hierarchical clustering method to generate pseudo labels and adopts a PK sampling in the training procedure. MMCL (Wang and Zhang 2020) and SSL (Lin et al. 2020) predict pseudo multi-labels by leveraging similarity computation and cycle consistency. After which they train the model as a multi-classification problem. SpCL (Ge et al. 2020) proposes a novel self-paced contrastive learning framework.

Unsupervised domain adaptation for person Re-ID

Recent studies on unsupervised domain adaptive person Re-ID could be mainly divided into two categories, one is the distribution aligning (Deng et al. 2018; Wei et al. 2018; Zhong et al. 2018a) , and the other is the clustering-based method(Song et al. 2020; Fu et al. 2019; Zhang et al. 2019; Yang et al. 2020; Ge, Chen, and Li 2020). For the distribution aligning, most methods use GAN-based methods to minimize the distance between the source and target domains. For example, SPGAN (Deng et al. 2018), and PT-GAN (Wei et al. 2018) use CycleGAN (Zhu et al. 2017) or StarGAN (Choi et al. 2018) to transfer images from the source domain style to the target domain style and train the model with the source domain identity labels. Although these methods try to align the distribution between the source domain and target domain, they can't fully explore the target domain images relationship and thus get unsatisfying performance.

Recently, clustering-based methods are widely used in domain adaptive person Re-ID. UDAP (Song et al. 2020) first utilizes the cluster method to generate pseudo labels. SSG (Fu et al. 2019) uses the global body and local parts to exploit the potential similarity in a clustering-guided approach. Although these clustering-based methods dominate this area, they still suffer the pseudo-label noise problem. Recently, more clustering-based methods (Zhang et al. 2019; Yang et al. 2020; Dai et al. 2021; Zheng et al. 2021a,b) are studying how to mitigate the influence of pseudo label noise. MMT (Ge, Chen, and Li 2020) proposes to softly refine the pseudo labels in the target domain by a deep mutual learning way. GLT(Zheng et al. 2021b) treat the pseudo label refinery problem as an transportation problem. UNRN(Zheng et al. 2021a) exploits the uncertainty to evaluate the reliability of

the pseudo-label of a sample. Different from the above work, in this paper, we propose a reliability exploration with self-ensemble learning method to alleviate the negative effect of noisy label in both online self-ensemble mutual learning and offline reliable sample selection .

Learning with noise

Recent studies on learning with noisy labels can broadly group to three categories: robust loss design (Wang et al. 2019), label correction (Lee et al. 2018) and re-weighting methods (Yang et al. 2020; Han et al. 2018). Robust loss design aims find a robustness function for noisy labels. Ghosh *et al.* (Ghosh, Kumar, and Sastry 2017) find that the mean absolute error loss is robust for noisy label. Wang *et al.* (Wang et al. 2019) later propose a symmetric cross entropy to avoid overfitting in CE loss. Label correction methods (Lee et al. 2018) try to estimate transition probabilities between noisy labels and true labels and try to correct noisy labels. However, these methods require additional clean data for training. Re-weighting methods reweight the loss to help handle the noise. Co-teaching (Han et al. 2018) propose to utilize two networks and select small loss samples to teach its peer network for further training. In this paper, our work leverages the consistency of different clusters to estimate sample reliability to select reliable samples and assign reliability weight on Re-ID loss to resists noisy labels produced by clustering.

Contrastive learning

Recently, contrastive learning is used in the field of unsupervised learning to learn a good image representation such as: MoCo(He et al. 2020), SimCLR(Chen et al. 2020) and BYOL(Grill et al. 2020), . Although these methods can learn a discriminative feature, they cannot well generalize to the person Re-ID task. Some methods introduce contrastive loss on person Re-ID, for example, SpCL(Ge et al. 2020) utilizes hybrid memory for contrastive learning. Different from SpCL, in this paper, we update the cluster-level feature in memory bank by the mean feature in the moving average model and supervised different branch feature learning by a multi-branch contrastive loss.

Our method

Overview

UDA in ReID aims at adapting the model trained on a labeled source domain dataset $D_s = \left\{ (x_i^s, y_i^s)|_{i=1}^{N_s} \right\}$ to an unlabeled target domain dataset $D_t = \left\{ x_i^t|_{i=1}^{N_t} \right\}$. Each image x_i^s in the source domain contains a corresponding ground-truth label y_i^s, and the target domain dataset contains N_t samples without their identity label. We aim to utilize the labeled data in D_s and the unlabeled data in D_t to learn a discriminative representations for the target dataset.

Fig. 2 shows the overall architecture of our proposed Reliability Exploration with Self-ensemble Learning (RESL) framework for UDA person Re-ID. We aim to relieve the negative influence of noisy pseudo labels in the cluster-based learning method. RESL firstly trains multi-branch model in the source domain dataset in a supervised manner. After that, the multi-branch model is adapted to the target dataset by iteratively training. Each branch can be viewed as an independent expert, in which different augmentations, drop blocks, and blocks of CNNs are used to help increase the diversity. Meanwhile, an instance-aware router is adopted to integrate different expert predictions. The ensemble predictions in the temporally average model are regarded as soft labels and are utilized to guide different branches in a mutual learning way. In each iteration, features from different branches of running model and the temporally averaged model are used to cluster and generate pseudo labels. The consistency in different cluster results is used to evaluate the reliability of instances. A cluster-level memory is adopted to guide multiple branches at feature level. In this way, the noise in pseudo labels can be effectively reduced.

Supervised training in source domains

The proposed RESL framework aims to transfer the knowledge of multi-branch experts from a labeled source domain to an unlabeled target domain. Therefore, we first train the Re-ID model on the labeled source domain. The RESL model consists of a stem CNN and K branches which is parametered with θ. The RESL model output K feature representations $f(x_{i,k}^s \mid \theta)$ and predicted probabilities $p(y_i^s \mid x_{i,k}^s, \theta)$, where $x_{i,k}^s$ is the ith sample's the kth augmentation image inputted into the kth branch. The cross-entropy loss for multiple branches can be defined as:

$$\mathcal{L}_{id}^s = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{k=1}^{K} \log p_j \left(y_i^s \mid x_{i,k}^s, \theta \right) \tag{1}$$

where $p(y_i^s \mid x_{i,k}^s, \theta)$ is the predicted probability of the sample x_i in the kth branch data flow. The softmax triplet loss can also be denoted as:

$$\mathcal{L}_{tri}^s = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{k=1}^{K} \log \frac{e^{\left\| f(x_{i,k}^s|\theta) - f(x_{i-,k}^s|\theta) \right\|}}{e^{\left\| f(x_{i,k}^s|\theta) - f(x_{i+,k}^s|\theta) \right\|} + e^{\left\| f(x_{i,k}^s|\theta) - f(x_{i-,k}^s|\theta) \right\|}} \tag{2}$$

where $f(x_{i,k}^s \mid \theta)$ is the feature for the source domain sample x_i^s in the kth branch. $x_{i+,k}^s$ and $x_{i-,k}^s$ mean the positive and negative samples for the ith sample. $\| \cdot \|$ represents the L_2 distance. The overall loss can be defined as :

$$\mathcal{L}^s = \mathcal{L}_{id}^s + \mathcal{L}_{tri}^s \tag{3}$$

With the K branches architecture, the model can produce K diverse features and predictions. The consistency between different features and the ensemble prediction can be used to enhance the model training.

Unsupervised training in target domain

As shown in Fig 2, the RESL framework contains four components: style augmentation, reliable sample selection, self-ensemble learning, and contrastive learning. After adaptation training, only the temporally averaged model is used during the inference stage.

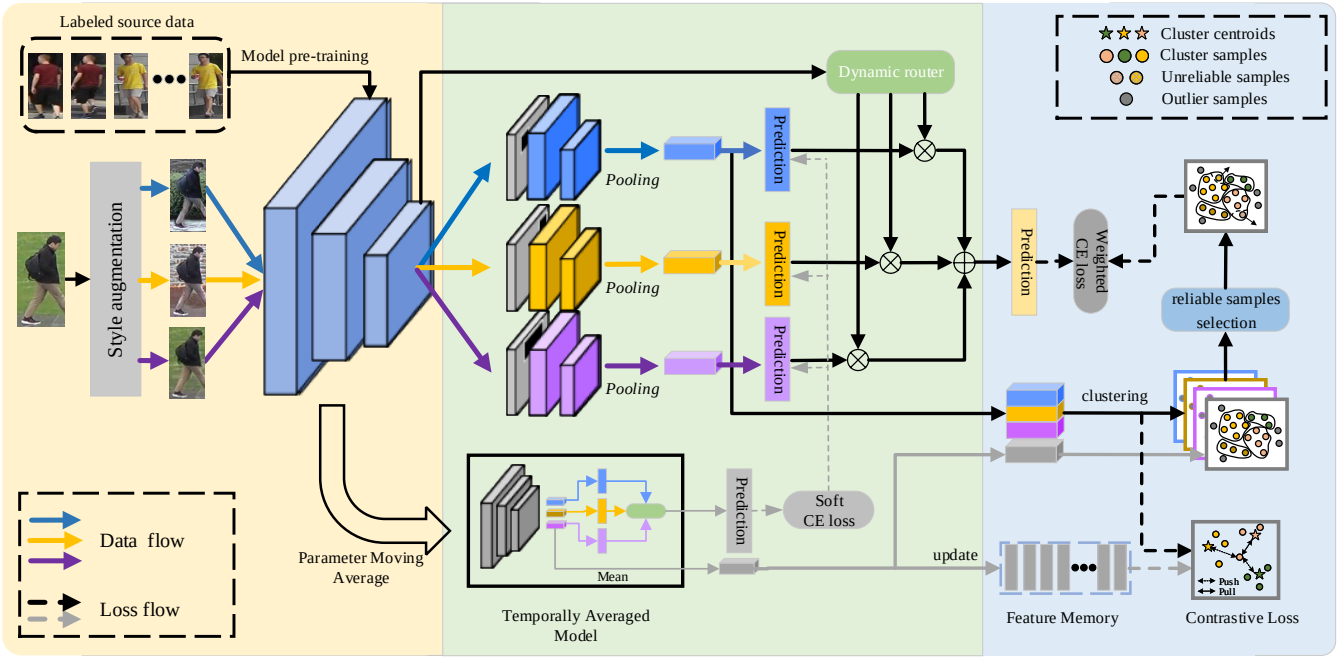

Figure 2: The framework of the proposed RESL consists of four components: data augmentation, self-ensemble learning module, reliable sample selection and contrastive learning module. The reliable sample selection utilizes on the consistency of different clusters to evaluate the reliability of identity samples and removes the unreliable ones. In the self-ensemble learning stage, a dynamic router is used to extract the weight to ensemble different branch predictions, and the ensemble predictions in the temporally average model are used as the soft labels to supervise the model learning. A contrastive loss is used on the cluster-level feature memory bank to enhance the feature discrimination.

Data augmentation In our method, data augmentation is critical to enhancing the diversity of different branches. In addition to traditional augmentation methods like Random Erasing (Zhong et al. 2020a), cropping, and flipping, We utilize Generative Adversarial Network (GAN) to generate additional data with different camera styles. By this means, we can obtain more diverse samples under various camera views, illumination, and background while preserving the original identities. We consider each camera as a style and utilize CycleGAN (Zhu et al. 2017) to train a camera-style transfer model (Zhong et al. 2018b) while preserving the original image identities. For an image x_i^t with its pseudo label y_i', we generate $C_t - 1$ images with different camera styles, where C_t is the camera number in the target domain. In each training iteration, we randomly select K augmented images for training and send them into different branches to equip different experts with diverse knowledge.

Reliability estimation with cluster consistency In order to identify unreliable samples assigned with noisy pseudo labels, we propose a reliability evaluation method by calculating the consistency of different clusters' results. Different branch features in the running model and the mean feature in the temporally averaged model are used to cluster the samples in the target domain separately. It should be noted that in our method, the temporally averaged model is designed to generate more stable features and predictions. The parameters of the temporally average model at the iteration T are denoted as $E^{(T)}[\theta]$, which can be calculated as:

$$E^{(T)}[\theta] = \begin{cases} \theta_t, & \text{if } t = 0, \\ \alpha E^{(T-1)}[\theta] + (1-\alpha)\theta_t, & \text{otherwise} \end{cases} \tag{4}$$

where $\alpha \in [0, 1]$ is the moving momentum, $E^{(T-1)}$ is the parameter of the temporally average model in the previous iteration $(T-1)$.

Let $f(x_{i,k}^t|\theta)$ represent the kth branch feature in the running model, and $f(x_i^t|E[\theta])$ means the ensemble feature in the temporally average model, which can be calculated as:

$$\overline{f}(x_i^t|E[\theta]) = \frac{1}{K}\sum_{k=1}^{K} f(x_{i,k}^t|E[\theta]). \tag{5}$$

Then, we use the cluster algorithm (*e.g.*, DBSCAN (Ester et al. 1996b)) on these extracted features $\{f(x_{i,1}^t|\theta), ..., f(x_{i,k}^t|\theta), \overline{f}(x_i^t|E[\theta])\}$. Thus, for a given sample x_i^t, we can obtain $K+1$ cluster results through these features, defined as $\{\mathcal{I}(f(x_{i,1}^t|\theta), ..., \mathcal{I}(f(x_{i,k}^t|\theta), ..., \mathcal{I}(\overline{f}(x_i^t|E[\theta])\}$.

Since each cluster is identified as a distinct class, the clustering reliability would significantly influence the model training. Hence, a reliability estimation strategy is introduced, where a reliable sample is measured by the consistency of different cluster results. A sample can be considered highly inconsistent if its different cluster results share various neighbors. Therefore, the following metric is presented to measure the sample reliability, which is formulated as an

intersection-over-union (IOU) score:

$$\mathcal{R}\left(f_i^t\right)=\frac{\left|\mathcal{I}(f(x_{i,1}^t|\theta))\cap\mathcal{I}(f(x_{i,k}^t|\theta))\ldots\cap\mathcal{I}(f(x_i^t|E[\theta]))\right|}{\left|\mathcal{I}(f(x_{i,1}^t|\theta))\cup\mathcal{I}(f(x_{i,k}^t|\theta))\ldots\cup\mathcal{I}(f(x_i^t|E[\theta]))\right|}\in[0,1] \tag{6}$$

where $\mathcal{I}(f(x_{i,k}^t|\theta))$ is the cluster set containing the feature $f(x_{i,k}^t|\theta)$, which is the feature of the ith sample in the kth branch. And $\overline{f}(x_i^t|E[\theta])$ is the ensemble feature in the temporally average model. A larger $\mathcal{R}\left(f_i^t\right)$ indicates the sample has more consistent neighbors in different cluster results, which means more reliable.

As shown in Fig 1, more reliable samples have more overlapping neighbors in different clusters. Given the above metrics to measure the reliability of the sample, more reliable samples can be selected for model training. We set $\beta\in[0,1]$ as the reliability threshold, preserving the samples whose $\mathcal{R}>\beta$ and considering the remaining samples as outliers.

Self-ensemble learning As shown in Fig 2, our RESL model uses an instance-aware router to adaptively integrate the prediction of each expert branch to each input sample. Taking ResNet as an example, we equipped the K expert branches with different drop blocks, ResBlocks and Generalize Mean Pooling(GeM) after the stem CNNs. The dynamic router consists of a global average pooling and a fully connected layer to produce an ensemble weight $w\in\mathbb{R}^K$, where K is the number of branches. Then, we apply the weight to aggravate the prediction of different branches to get an ensemble prediction:

$$\overline{p}\left(y_i'|\,x_i^t,\theta\right)=\sum_{k=1}^{K}w^k p\left(y_i'\mid x_{i,k}^s,\theta\right) \tag{7}$$

where y_i' is the clustering-based label and $p(y_i'\mid x_{i,k}^s,\theta)$ is the prediction of the kth branch for the ith image. During the training stage, a reliability-weighted classification loss is adopted on the branch and ensemble predictions, which can be formulated as:

$$\mathcal{L}_{ce}=-\frac{1}{N_t}\sum_{i=1}^{N_t}R(f_i)\left(\sum_{k=1}^{K}\log p\left(y_i'\mid x_{i,k}^t,\theta\right)+\log\overline{p}\left(y_i'\mid x_i^t,\theta\right)\right) \tag{8}$$

in which $R(f_i)$ is the reliability of the sample x_i^t.

To aggregate diverse knowledge and reduce the negative influence of noisy labels in an online training way, we take the ensemble predictions in the temporally averaged model as soft labels.

Given a target-domain sample and its augmented images, the temporally average network encodes them into an ensemble prediction $\overline{p}\left(y_i'\mid x_i^t,E[\theta]\right)$. Then, the ensemble prediction supervises the current running model by a soft cross-entropy loss in a mutual learning way. It can be denoted as:

$$L_{sce}=-\sum_{i=1}^{N_t}\sum_{k=1}^{K}\overline{p}\left(y_i'\mid x_i^t,E[\theta]\right)\log p\left(y_i'\mid x_{i,k}^t,\theta\right) \tag{9}$$

where $\overline{p}\left(y_i'\mid x_i^t,E[\theta]\right)$ is the dynamic ensemble prediction in the temporally average model. During self-ensemble learning, each branch in the running model can be corrected in an online way by the ensemble soft labels. Since the data augmentation is applied, different branches also learn different knowledge from data, which further improves the ensemble prediction quality.

Contrastive learning A contrastive loss is adopted to enhance the feature discrimination. But different from the existing method (Ge et al. 2020), the cluster-level memory bank is updated by using the ensemble feature in the temporally averaged model and the contrastive loss is applied on the multi-branch features. Firstly, the memory bank can be initialized by:

$$\boldsymbol{c}_j=\frac{1}{|\mathcal{I}_j|}\sum_{x_i^t\in\mathcal{I}_j}\overline{f}(x_i^t|E[\theta]) \tag{10}$$

where $\overline{f}(x_i^t|E[\theta])$ is the ensemble feature in the temporally average network, and $\mathcal{I}_j$ is the jth refined cluster where x_i is located.

During each iteration, the centroid $\boldsymbol{c}_j$ can be updated by the following equation:

$$\boldsymbol{c}_j\leftarrow m\boldsymbol{c}_j+(1-m)\,\overline{f}(x_i^t|E[\theta]) \tag{11}$$

where m is the momentum factor that updates centroids. With centroids in the memory bank, the similarity between samples and classes can be measure by dot product. Therefore, the contrastive loss can be denoted as:

$$L_{con}=-\frac{1}{N_t}\sum_{i=1}^{N_t}\sum_{k=1}^{K}\log\frac{\exp\left(f(x_{i,k}^t|\theta)\cdot c^+\right)/\tau}{\sum_{j=1}^{J}\exp\left(f(x_{i,k}^t|\theta),c_j\right)/\tau} \tag{12}$$

where c^+ is the class feature the sample x_i^t belongs to, and τ is the temperature parameter.

Finally, the total loss for the target domain can be formulated as:

$$\mathcal{L}_{\text{target}}=\lambda_{ce}L_{ce}+\lambda_{sce}L_{sce}+\lambda_{con}L_{con} \tag{13}$$

where λ_{ce}, λ_{sce} and λ_{con} are weighting factors.

Experiments

Datasets and Evaluation Protocol

We evaluate our method on three large-scale Re-ID datasets: Market-1501 (Zheng et al. 2015), DukeMTMC-ReID (Ristani et al. 2016; Zheng, Zheng, and Yang 2017) and MSMT17 (Wei et al. 2018).

Market-1501 (Zheng et al. 2015) consists of 1501 identities with 32,668 images, which was captured by 6 different cameras. The training set contains 751 identities with 12,936 images. The test set includes 750 identities, where the query set contains 3,368 images and the gallery set contains 19,732 images.

DukeMTMC-ReID (Ristani et al. 2016) is a sub-dataset of DukeMTMC (Zheng, Zheng, and Yang 2017), which includes 36,411 images with 1812 identities. All images are collected from 8 high-definition cameras. In addition, 16,522 training images, 2,228 queries images, and 17,661 gallery images are in the dataset.

Table 1: Performance comparison of the proposed method and state-of-the-art methods for domain adaptation on DukeMTMC-ReID, Market1501 and MSMT17 datasets.

Methods	Reference	DukeMTMC-ReID to Market1501				Market1501 to DukeMTMC-ReID			
		mAP	R1	R5	R10	mAP	R1	R5	R10
PUL (Fan et al. 2018)	TOMM 2018	20.5	45.5	60.7	66.7	16.4	30.0	43.4	48.5
SPGAN+LMP (Deng et al. 2018)	CVPR 2018	26.7	57.7	75.8	82.4	26.2	46.4	62.3	68.0
CamStyle (Zhong et al. 2018b)	TIP 2018	27.4	58.8	78.2	84.3	25.1	48.4	62.5	68.9
ECN (Zhong et al. 2019)	CVPR 2019	43.0	75.1	87.6	91.6	40.4	63.3	75.8	80.4
PDA-Net (Li et al. 2019)	ICCV 2019	47.6	75.2	86.3	90.2	45.1	63.2	77.0	82.5
UDAP (Song et al. 2020)	PR 2020	53.7	75.8	89.5	93.2	49.0	68.4	80.1	83.5
SSG (Fu et al. 2019)	ICCV 2019	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
MMCL (Wang and Zhang 2020)	CVPR 2020	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0
ACT (Yang et al. 2020)	AAAI 2020	60.6	80.5	-	-	54.5	72.4	-	-
ECN-GPP (Zhong et al. 2020b)	TPAMI 2020	63.8	84.1	92.8	95.4	54.4	74.0	83.7	87.4
AD-Cluster (Zhai et al. 2020a)	CVPR 2020	68.3	86.7	94.4	96.5	54.1	72.6	82.5	85.5
MMT (Ge, Chen, and Li 2020)	ICLR 2020	71.2	87.7	94.9	96.9	65.1	78.0	88.8	92.5
MEB-Net (Zhai et al. 2020b)	ECCV 2020	76.0	89.9	96.0	97.5	66.1	79.6	88.3	92.2
SpCL (Ge et al. 2020)	NIPS 2020	77.5	89.7	96.1	97.6	68.8	82.9	90.1	92.5
Dual-Refinement (Dai et al. 2021)	TIP 2021	78.0	90.9	96.4	97.7	67.7	82.1	90.1	92.5
UNRN (Zheng et al. 2021a)	AAAI 2021	78.1	91.9	96.1	97.8	69.1	82.0	90.7	93.5
GLT (Zheng et al. 2021b)	CVPR 2021	79.5	92.2	96.5	97.8	69.2	82.0	90.2	92.8
Ours	**this paper**	**83.1**	**93.2**	**96.8**	**98.0**	**72.3**	**83.9**	**91.7**	**93.6**

Methods	Reference	Market-1501 to MSMT17				DukeMTMC-ReID to MSMT17			
		mAP	R1	R5	R10	mAP	R1	R5	R10
ECN (Zhong et al. 2019)	CVPR 2019	10.2	30.2	41.5	46.8	8.5	25.3	36.3	42.1
SSG (Fu et al. 2019)	ICCV 2019	13.3	32.2	-	51.2	13.2	31.6	-	49.6
MMCL (Wang and Zhang 2020)	CVPR 2020	16.0	42.5	55.9	61.5	15.2	40.4	53.1	58.7
ECN-GPP (Zhong et al. 2020b)	TPAMI 2020	16.2	43.6	54.3	58.9	15.1	40.8	51.8	56.7
MMT (Ge, Chen, and Li 2020)	ICLR 2020	23.3	50.1	63.9	69.8	22.9	49.2	63.1	68.8
JVTC+ (Li and Zhang 2020)	ECCV 2020	25.1	48.6	65.3	68.2	27.5	52.9	70.5	75.9
SpCL (Ge et al. 2020)	NeurIPS 2020	26.8	53.7	65.0	69.8	26.5	53.1	65.8	70.5
Dual-Refinement (Dai et al. 2021)	TIP 2021	25.1	53.3	66.1	71.5	26.9	55.0	68.4	73.2
UNRN (Zheng et al. 2021a)	AAAI 2021	25.3	52.4	64.7	69.7	26.2	54.9	67.3	70.6
GLT (Zheng et al. 2021b)	CVPR 2021	26.5	56.6	67.5	72.0	27.7	59.5	70.1	74.2
Ours	**this paper**	**33.6**	**64.8**	**74.6**	**79.6**	**34.2**	**65.2**	**74.6**	**80.1**

MSMT17 (Wei et al. 2018) is a large-scale dataset, containing 126,441 images of 4,101 identities. The training set contains 1,041 identities and the test set contains 3,060 identities.

Cumulative Matching Characteristic (CMC) and Mean Average Precision (mAP) are used to evaluate the model. All experiment results are under the single-query setting, and no post-processing (Zhong et al. 2017) is applied.

Implementation Details

We use ResNet50 pre-trained on ImageNet(Deng et al. 2009) as our stem CNNs. We add dropblocks, ResBlock, and Generalize-Mean pooling(GeM) after the 4th layer of ResNet50, which compose the structure of the branch. In addition to random erasing, flipping and cropping, we augmented the images with different camera styles using the style transfer model. The number of K is set to 3, and each image has three augmented samples in each training batch. The batch size for the source domain and target domain are both set as 128, containing 16 identities. We utilize the DBSCAN clustering algorithm, and the Jaccard distance with k-reciprocal nearest neighbor is used as the distance metric. The eps in DBSACN is set to 0.6. SGD optimizer is adopted for model optimization. The initial learning rate is set to 0.00035, and is divided by 10 at the 40th and 60th epoch, in a total 70 epochs. The reliability threshold β is set to 0.2. the momentum factor α and m in Eq.(4) and Eq.(11) are set to 0.999 and 0.2. Our model is implemented on PyTorch (Paszke et al. 2019) platform.

Comparison with state-of-the-arts

As shown in Table 1, Our method obtains the performance of 83.1% on mAP and 93.2% on rank-1 when transferring DukeMTMC-ReID to Market1501. RESL outperforms the best memory-based method SPCL by 5.6% and 3.5% on mAP and rank-1 accuracy. Moreover, the RESL method outperforms MMT by 11.9% and 5.5% and MEB-Net by 6.1% and 3.5% on mAP and rank-1, which both utilize multiple models and mean net for mutual learning. For DukeMTMC-ReID to Market-1501, our method also gains the best UDA performance and outperforms the second method GLT by 3.6% and 1.0% on mAP and rank-1.

MSMT17 is a large dataset, consisting 126,441 images and 4,101 identities, which is a more difficult job. And our method still gets satisfying results on this challenging dataset. As shown in TABLE 1, our method achieve 33.6% mAP and 64.8% rank-1 when transfer DukeMTMC-ReID to MSMT17, which is 7.1% and 8.2% higher than GLT on mAP and rank-1. When considering DukeMTMC-ReID as the source dataset, we get 34.2% and 65.2% on mAP and rank-1, which also outperform other methods. The satisfying results on this challenging dataset verify our method's

Table 2: Ablation study on the effectiveness of components in RESL method. Contrastive loss (CT): utilizing contrastive loss for model optimizing. Temporally average model (TM): utilizing temporally average model as mean teacher model. Reliability weighted loss(WL): Incorporate the reliability to the re-id loss. Ensemble learning(EL): utilizing muti-branch for ensemble learning. Reliable Sample Selection(RSS): Selecting reliable sample for training, which is performed together with EL. Single Branch: only one branch feature is used when inferencing.

Methods	Duke to Market		Market to Duke	
	mAP	R1	mAP	R1
Fully supervised	84.6	94.3	76.2	87.1
Direct Transfer	28.6	58.0	27.6	44.5
Baseline	69.1	87.7	61.2	75.9
Base+TM+CT	79.0	91.7	69.2	81.9
Base+TM+CT+WL	79.5	92.1	70.2	82.3
Base+TM+CT+EL	81.3	92.5	70.6	83.1
Base+TM+CT+RSS	82.1	92.8	71.2	83.3
Single Branch	80.7	92.6	71.0	82.6
All	**83.1**	**93.2**	**72.3**	**83.9**

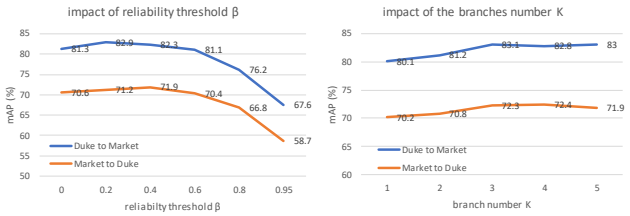

Figure 3: Performance comparison with different reliability threshold β and different branch number K.

effectiveness.

Ablation Studies

In this section, we perform ablation experiments on different components of our method to evaluate their effectiveness.

Table 3: The computational complexity comparison. Mem denotes the training memory when bs is 64.

Methods	DukeMTMC-ReID to Market1501		
	R1(mAP)	Params	Mem
MMT	87.7(71.2)	23.51M	14,735M
MEB-Net	90.3(76.7)	23.51M	16,856M
ours(Single)	92.1(80.5)	**23.51M**	-
ours	**92.8(82.0)**	32.94M	12,539M

1) Comparisons between supervised learning: In Table 2, we compare the performance with supervised learning, direct transfer, baseline model, and RESL. The fully supervised learning utilizes the ground-truth label to train the model and thus get the best performance. When directly transfer the model from Market-1501 to DukeMTMC-ReID, the performance of mAP drops from 83.5% to 28.6%, which means there is a large domain gap between the two datasets. The baseline model utilizes only classification loss and triplet loss, which is the same as the baseline described in MMT(Ge, Chen, and Li 2020). Our method improves the mAP from 69.1% to 83.1% compared with the baseline. And even use a single branch in the inference stage, our method can also achieve 80.7% mAP and 92.6% rank-1, which is superior to most state-of-the-art methods.

2) Effectiveness of self-ensemble learning: In Table 2, we evaluate the effectiveness of self-ensemble, which uses an ensemble-based soft CE loss to supervise the model learning. When adding the ensemble learning, the performance outperforms the baseline+TM+CT by 1.2% on mAP and 1.0% on rank-1. This is because self-ensemble learning can reduce the weight of unreliable samples and help the model learn more useful information.

3) Effectiveness of the reliable sample selection: In Table 2, we evaluate the effectiveness of reliable sample selection (RSS). With the reliable sample selection, our method further improves the mAP and rank-1 by 2.6% mAP and 1.7% rank-1 on the Duke to Market, and 2.0% and 1.4% on Market to Duke. This shows that reliable sample selection can further relieve the negative effect of the noisy pseudo label in the target dataset and gain better performance.

4) Analysis the branch number K: K is the branch number in the RESL framework. As illustrated in Fig 3, we investigate the effect of different values of K by changing its value from 1 to 5. And it can be observed that with the more branch participating in, the performance gradually increasing. But the performance stop growing when K is larger than 3. This is because more diverse knowledge and features can be captured as the branch number increasing. On the other hand, too many branches also can't promote model training.

5) The impact of reliability threshold β: β is used to select reliable samples for model training. In Fig 3, we investigate the effect of different value α. And we can find that when $\beta < 0.2$ or $\beta > 0.4$, the performance decreases. This is because, with a small β, samples with noisy pseudo label can not be found. But when β is too large, less sample can be used for training, which also hinders the performance.

6) Computational complexity analysis : The quantitative comparisons in Table 3 show our method requires less training memory and gains performance improvement under acceptable parameter growth. In addition, our **single** branch achieves superior performance than other methods using the same amount of parameters.

Conclusion

In this paper, we propose a Reliability Exploration with self-Ensemble Learning (RESL) framework to handle the noisy pseudo-label problem in clustering-based UDA person ReID in both an online self-ensemble learning and an offline reliable sample selection way. First, a novel multiple branch scheme is proposed, and its temporally average model is used as a teacher model, where the ensemble predictions are treated as soft labels. Second, a reliability estimation method is proposed, which utilizes the consistency of different clusters' results to select reliable samples. These two techniques significantly alleviate the negative influence of wrong/noisy labels during the adaptation. And our method achieves state-of-the-art performance on benchmark datasets.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 61972169, in part by the National key research and development program of China(2019QY(Y)0202), in part by the Major Scientific and Technological Project of Hubei Province (2019AAA051) and the Research Programme on Applied Fundamentals and Frontier Technologies of Wuhan(2020010601012182).

References

Bak, S.; Carr, P.; and Lalonde, J.-F. 2018. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 189–205.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Dai, Y.; Liu, J.; Bai, Y.; Tong, Z.; and Duan, L.-Y. 2021. Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. *IEEE Transactions on Image Processing*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 994–1003.

Ding, Y.; Fan, H.; Xu, M.; and Yang, Y. 2020. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1): 1–19.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996a. Density-based spatial clustering of applications with noise. In *Int. Conf. Knowledge Discovery and Data Mining*, volume 240, 6.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996b. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, 226–231.

Fan, H.; Zheng, L.; Yan, C.; and Yang, Y. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4): 1–18.

Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; and Huang, T. S. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6112–6121.

Ge, Y.; Chen, D.; and Li, H. 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; and Li, H. 2020. Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID. In *Advances in Neural Information Processing Systems*.

Ghosh, A.; Kumar, H.; and Sastry, P. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Lee, K.-H.; He, X.; Zhang, L.; and Yang, L. 2018. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5447–5456.

Li, J.; and Zhang, S. 2020. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *European Conference on Computer Vision*, 483–499. Springer.

Li, Y.-J.; Lin, C.-S.; Lin, Y.-B.; and Wang, Y.-C. F. 2019. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7919–7929.

Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; and Yang, Y. 2019a. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8738–8745.

Lin, Y.; Xie, L.; Wu, Y.; Yan, C.; and Tian, Q. 2020. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3390–3399.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; and Yang, Y. 2019b. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95: 151–161.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, 17–35. Springer.

Shi, Y.; Ling, H.; Wu, L.; Shen, J.; and Li, P. 2020a. Learning refined attribute-aligned network with attribute selection for person re-identification. *Neurocomputing*, 22071.

Shi, Y.; Ling, H.; Wu, L.; Zhang, B.; and Li, P. 2021. Attribute disentanglement and registration for occluded person re-identification. *Neurocomputing*.

Shi, Y.; Wei, Z.; Ling, H.; Wang, Z.; Shen, J.; and Li, P. 2020b. Person Retrieval in Surveillance Videos via Deep Attribute Mining and Reasoning. *IEEE Transactions on Multimedia*.

Shi, Y.; Wei, Z.; Ling, H.; Wang, Z.; Zhu, P.; Shen, J.; and Li, P. 2020c. Adaptive and Robust Partition Learning for Person Retrieval with Policy Gradient. *IEEE Transactions on Multimedia*.

Song, L.; Wang, C.; Zhang, L.; Du, B.; Zhang, Q.; Huang, C.; and Wang, X. 2020. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 102: 107173.

Wang, D.; and Zhang, S. 2020. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10981–10990.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 322–330.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 79–88.

Yang, F.; Li, K.; Zhong, Z.; Luo, Z.; Sun, X.; Cheng, H.; Guo, X.; Huang, F.; Ji, R.; and Li, S. 2020. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12597–12604.

Zeng, K.; Ning, M.; Wang, Y.; and Guo, Y. 2020. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13657–13665.

Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; and Tian, Y. 2020a. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9021–9030.

Zhai, Y.; Ye, Q.; Lu, S.; Jia, M.; Ji, R.; and Tian, Y. 2020b. Multiple expert brainstorming for domain adaptive person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 594–611. Springer.

Zhang, X.; Cao, J.; Shen, C.; and You, M. 2019. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8222–8231.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4320–4328.

Zheng, K.; Lan, C.; Zeng, W.; Zhang, Z.; and Zha, Z.-J. 2021a. Exploiting Sample Uncertainty for Domain Adaptive Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3538–3546.

Zheng, K.; Liu, W.; He, L.; Mei, T.; Luo, J.; and Zha, Z.-J. 2021b. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5310–5319.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2138–2147.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 3754–3762.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1318–1327.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020a. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13001–13008.

Zhong, Z.; Zheng, L.; Li, S.; and Yang, Y. 2018a. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–188.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 598–607.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2020b. Learning to adapt invariance in memory for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*.

Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; and Yang, Y. 2018b. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28(3): 1176–1190.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.