Enhancing Diversity in Teacher-Student Networks via Asymmetric branches for Unsupervised Person Re-identification

Anonymous WACV submission

Paper ID 67

Abstract

The objective of unsupervised person re-identification (Re-ID) is to learn discriminative features without labor-intensive identity annotations. State-of-the-art unsupervised Re-ID methods assign pseudo labels to unlabeled images in the target domain and learn from these noisy pseudo labels. Recently introduced Mean Teacher Model is a promising way to mitigate the label noise. However, during the training, self-ensembled teacher-student networks quickly converge to a consensus which leads to a local minimum. We explore the possibility of using an asymmetric structure inside neural network to address this problem. First, asymmetric branches are proposed to extract features in different manners, which enhances the feature diversity in appearance signatures. Then, our proposed cross-branch supervision allows one branch to get supervision from the other branch, which transfers distinct knowledge and enhances the weight diversity between teacher and student networks. Extensive experiments show that our proposed method can significantly surpass the performance of previous work on both unsupervised domain adaptation and fully unsupervised Re-ID tasks.

1. Introduction

Person re-identification (Re-ID) targets at retrieving a person of interest across non-overlapping cameras. Since there are domain gaps resulting from illumination condition, camera property and view-point variation, a Re-ID model trained on a source domain usually shows a huge performance drop on other domains.

Unsupervised Domain Adaptation (UDA) targets at shifting the model trained from a source domain with identity annotation to a target domain via learning from unlabeled target images. In the real world, unlabeled images in a target domain can be easily recorded, which is almost labor-free. It is intuitive to use these images to adapt a pretrained Re-ID model to the desired domain. Fully unsupervised Re-ID further minimises the supervision by removing pre-training on the labelled source domain.

State-of-the-art UDA Person Re-ID methods [9, 28] and unsupervised methods [18] assign pseudo labels to unlabeled target images. The generated pseudo labels are generally very noisy. The noise is mainly from several inevitable factors, such as the strong domain gaps and the imperfection of clustering. In this way, an unsupervised Re-ID problem is naturally transferred into Generating pseudo labels and Learning from noisy labels problems, which is similar to how unlabeled samples are used in Semi-supervised learning.

To generate pseudo labels, the most intuitive way is to use a clustering algorithm, which gives a good starting point for clustering based UDA Re-ID [30, 7]. Recently, Ge *et al*. [9] propose to add a Mean Teacher [24] model as online soft pseudo label generator, which effectively reduces the error amplification during the training with noisy labels. In this paper, we also use both clustering-based hard labels and teacher-based soft labels in our baseline. We use a density based clustering (*i.e.*, DBSCAN [6]) and dynamically change the dimension of classifier, which surpasses the performance of K-Means++ [1] with dimension-fixed classifier used in [9].

To handle noisy labels, one of the most popular approaches is to train paired networks so that each network helps to correct its peer, e.g., two-student networks in Co-teaching [10] and two-teacher-two-student networks in MMT [9]. However, these paired models with identical structure are prone to converge to each other and get stuck in a local minimum. There are several attempts to alleviate this problem, such as Co-teaching+ [29], ACT [28] and MMT [9]. These attempts of keeping divergence between paired models are mainly based on either different training sample selection [29, 28] or different initialization and data augmentation[9]. In this paper, we propose a strong alternative by designing asymmetric neural network structure in the Mean Teacher Model. We use two independent branches with different depth and global pooling methods as last layers of a neural network. Features extracted from both branches are concatenated as the appearance signature, which enhances the feature diversity in the appearance signature and allows to get better clustering-based hard labels. Soft pseudo labels generated by the teacher network are used to supervise the student network in a cross-branch manner, which enhances the divergence between paired teacher-student networks. Our proposed decoupling method does not rely on different source domain initializations, which makes it more effective in the fully unsupervised scenario where the source domain is not available.

In summary, our contributions are:

1. We propose to enhance the feature diversity inside person Re-ID appearance signatures by splitting last layers of a backbone network into two asymmetric branches, which increases the quality of clustering-based hard labels.

2. We propose a novel decoupling method where asymmetric branches get cross-branch supervision, which avoids weights in paired teacher-student networks converging to each other and increases the quality of teacher-based soft labels.

3. Extensive experiments and ablation study are conducted to validate the effectiveness of each proposed component and the whole framework.

2. Related Work

Unsupervised domain adaptive Re-ID. Recent unsupervised cross-domain Re-ID methods can be roughly categorized into distribution alignment and pseudo label based adaptation. The objective of distribution alignment is to learn domain invariant features. Several attempts [25, 16] leverage semantic attributes to align the feature distribution in the latent space. However, these approaches strongly rely on extra attribute annotation, which require extra labor. Another possibility is to align the feature distribution by transferring labeled source domain images into the style of target domain with generative adversarial networks [26, 34, 3]. Style transferred images are usually combined with pseudo label based adaptation to get a better performance. Pseudo label based adaptation is a more straightforward approach for unsupervised cross-domain Re-ID, which directly assigns pseudo labels to unlabelled target images and allows to fine-tune a pre-trained model in a supervised manner. Clustering algorithms are widely used in previous unsupervised cross-domain Re-ID methods. UDAP [23] provides a good analysis on clustering based adaptation and use a k-reciprocal encoding [32] to improve the quality of clusters. PCB-PAST [30] simultaneously learns from a ranking-based and clustering-based triplet losses. SSG [7] assigns clustering-based pseudo labels to both global and local features. To mitigate the clustering-based label noise, researchers borrow ideas from how unlabeled data is used in Semi-supervised learning and Learning from noisy labels. ENC [35] uses an exemplar memory to save averaged features to assign soft labels. ACT [28] splits the training data into inliers/outliers to enhance the divergence of paired networks in Co-teaching [10]. MMT [9] adopts two student and two Mean Teacher networks. Two students are initialized differently from source pre-training in order to enhance the divergence of paired teacher-student networks. Each mean teacher network provides soft labels to supervise peer student network. However, despite different initializations at the beginning of adaptation, the decoupling is not encouraged enough during the training. We directly use asymmetric neural network structure inside teacher-student networks, which encourages the decoupling at all epochs.

Fully unsupervised Re-ID. Recently, several fully unsupervised Re-ID methods are proposed to further minimize the supervision, which does not require any Re-ID annotation. A bottom-up clustering framework is proposed in BUC [17], which trains a network based on the clustering-based pseudo labels in an iterative way. [18] replaces clustering-based pseudo labels with similarity-based softened labels. Different to image-based unsupervised Re-ID, [27] learns tacklet information with clustering-based pseudo labels. In our proposed method, both hard and softened pseudo labels are used. Asymmetric structure is proposed to enhance the diversity during the training process to increase the quality of pseudo labels, which helps us to outperform state-of-the-art methods.

Teacher-Student Network for Semi-Supervised Learning. Unsupervised domain adaptation can be regarded to some extent as Semi-Supervised Learning (SSL), since both of them utilize labeled data (source domain for UDA) and large amount of unlabeled data (target doamin for UDA). A teacher-student structure is commonly used in SSL. This structure allows student network to gradually exploit unlabeled data under consistency constraints. In Π model and Temporal ensembling [15], the student learns from either samples forwarded twice with different noise or exponential moving averaged (EMA) predictions under consistency constraints. Instead of EMA predictions, Mean-teacher model [24] use directly the EMA weights from the student to supervise the student under a consistency constraint. Authors of Dual student [14] point out that the Mean Teacher converging to student along with training (coupling problem) prevents the teacher-student from exploiting more meaningful information from data. Inspired by Deep Co-training [21], they propose to train two independent students on stable samples which have same predictions and enough large feature difference. However, in unsupervised cross-domain Re-ID, labeled source domain and unlabeled target

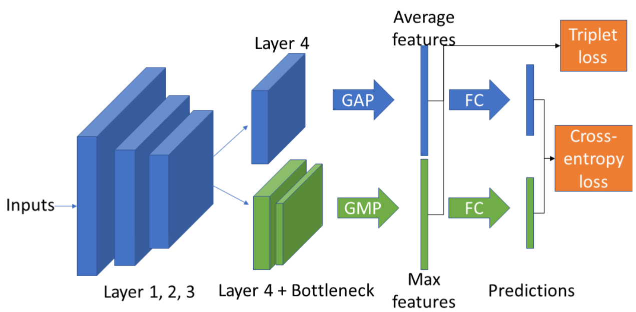

Figure 1. Source domain pre-training for asymmetric branched network. One ResNet bottleneck block corresponds to three convolutional layers. For UDA setting, inputs are labelled images from source training set.

domain do not share the same identity classes, which makes traditional close-set SSL methods hard to use.

3. Proposed Method

3.1. Overview

Given two datasets: one labeled source dataset D_s and one unlabeled target dataset D_t, the objective of UDA is to adapt a source pretained model M_{pre} to the target dataset with unlabeled target data. To achieve this goal, we propose a two-staged adaptation approach based on Mean Teacher Model. We focus on the coupling problem (teacher and student converge to each other) existing inside the original Mean Teacher. Asymmetric branches and cross-branch supervision are proposed in this paper to address this problem and to enhance the diversity in the network, which show great effectiveness for UDA Re-ID.

3.2. Asymmetric branches

A multi-branch structure is widely used in the fully supervised Re-ID methods, especially in global-local feature based methods [8, 4, 2]. Such structure keeps independence between branches, which makes features extracted from different branches diversified. In the unsupervised Re-ID, we conduct clustering on appearance signatures to generate pseudo labels. The quality of pseudo labels is strongly depended on the quality of appearance signatures. We want to extract distinct meaningful features from different branches. Thus, we duplicate last layers of a backbone network and make them different in the structure, which we call Asymmetric Branches.

Asymmetric branches are illustrated in Figure 1. For a ResNet-based [11] backbone, the layer 4 is duplicated. The first branch is kept unchanged as the one used in the original backbone: 3 bottlenecks and global average pooling (GAP). The second branch is composed of 4 bottlenecks and global max pooling (GMP). The GAP perceives global information, while the GMP focuses on the most discriminative information (most distinguishable identity information, such as a red bag or a yellow t-shirt). Asymmetric branches improve appearance signature quality by enhancing the feature diversity, which is validated by source pre-training performance boost in Table 3 as well as examples in Figure 5. They further improve the quality of pseudo labels during the adaptation, which is validated by target adaptation performance in Table 3.

3.3. Asymmetric Branched Mean Teaching

We call our proposed adaptation method Asymmetric Branched Mean Teaching (ABMT). Our proposed ABMT contains two stages: Source pre-training and Target adaptation.

3.3.1 Source domain supervised pre-training

In the first stage, we train a network in the fully supervised way on the source domain. Thanks to this stage, the model used for adaptation obtains a basic Re-ID capacity, which helps to alleviate pseudo label noise. Given a source sample x_i^s and its ground truth identity y_i', the network (with weight θ) encodes x_i^s into average $F_a(x_i^s|\theta)$ and max features $F_m(x_i^s|\theta)$ and then gets two predictions $P_a(x_i^s|\theta)$ and $P_m(x_i^s|\theta)$. Cross-entropy and batch hard triplet [12] losses are used in this stage as shown in Figure 1.

$$L_{ce}(y_i, y_i') = -\sum_i y_i' \log(y_i) \tag{1}$$

$$\begin{aligned} L_{tri}(\mathbf{a_i}, \mathbf{p_i}, \mathbf{n_j}) = \sum_{i=1}^{P}\sum_{a=1}^{K}[&\max_{p=1,\dots,K} \|\mathbf{a_i} - \mathbf{p_i}\|_2 \\ &- \min_{\substack{n=1,\dots,K \\ j=1,\dots,P \\ j\neq i}} \|\mathbf{a_i} - \mathbf{n_j}\|_2 + \alpha]_+ \end{aligned} \tag{2}$$

where $\|\mathbf{a_i} - \mathbf{p_i}\|_2$ is Euclidean distance between anchor feature vector $\mathbf{a_i}$ and positive feature vector $\mathbf{p_i}$, while $\|\mathbf{a_i} - \mathbf{n_j}\|_2$ is Euclidean distance between anchor feature vector $\mathbf{a_i}$ and negative feature vector $\mathbf{n_i}$.

The whole network is trained with a combination of both losses:

$$\begin{aligned} L_{scr} =& \lambda_{ce}^s L_{ce}(P_a(x_i^s|\theta), y_i') + \lambda_{ce}^s L_{ce}(P_m(x_i^s|\theta), y_i') \\ &+ \lambda_{tri}^s L_{tri}(P_a(x_i^s|\theta), P_a(x_p^s|\theta), P_a(x_n^s|\theta)) \\ &+ \lambda_{tri}^s L_{tri}(P_m(x_i^s|\theta), P_m(x_p^s|\theta), P_m(x_n^s|\theta)) \end{aligned} \tag{3}$$

3.3.2 Target domain unsupervised adaptation

The adaptation procedure is illustrated in Figure 2. It contains two components: Clustering-based hard label generation and Cross-branch teacher-based soft label training. After adaptation, only teacher network is used during the inference.

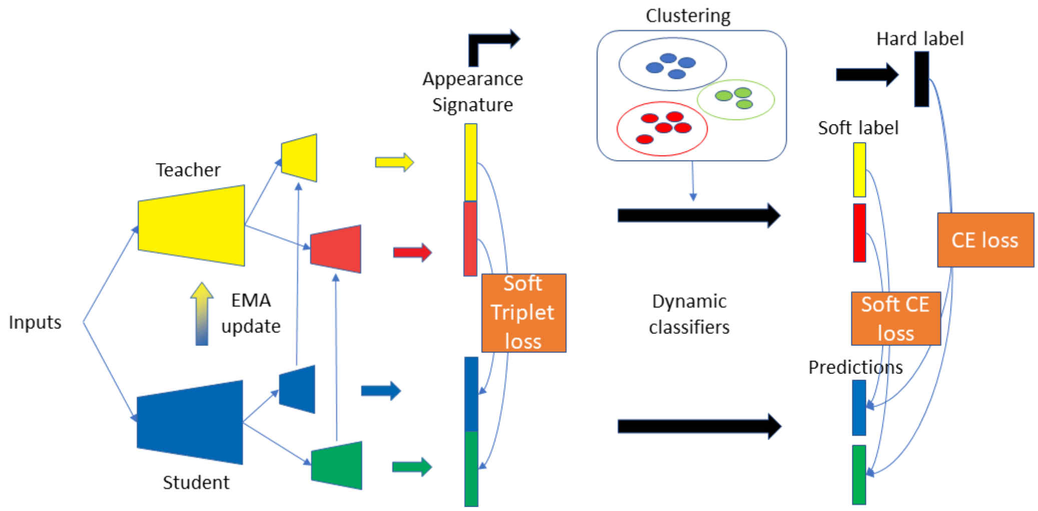

Figure 2. ABMT adaptation. For UDA setting, inputs are training set images from both source and target domains. For fully unsupervised setting, inputs are unlabeled images from target training set.

Clustering-based hard label generation. In previous UDA Re-ID methods, distance-based K-Means [9] and density-based clustering DBSCAN [28, 23] are main approaches to generate pseudo labels. In the real world, it is hard to know the class number in the target domain, which makes K-Means unpractical.

We follow the state-of-the-art density-based clustering method in [23]. To adapt it to our proposed asymmetric branches, we concatenate the average and max features from asymmetric branches in the teacher network as appearance signatures. Images belonging to the same identity should have same nearest neighbors in the feature space. Distance metric for DBSCAN are obtained by k-reciprocal re-ranking encoding [32] between target domain and source domain samples.

A density-based clustering generates unfixed cluster numbers at different epochs, which means old classifiers from last epoch can not be reused after a new clustering. Thus, we simply create new classifiers depending on the number of clusters at the beginning of each epoch. We take normalized mean features of each cluster from average branch to initialize the average branch classifiers and similarly those from max branch to initialize the max branch classifiers. We call them Dynamic Classifiers. With the help of Dynamic Classifiers, the student is trained on cluster components (outliers are discarded) with cross-entropy loss:

$$L_{ce} = -\sum_i (y_i' \log(P_m(x_i^t|\theta))) - \sum_i (y_i' \log(P_a(x_i^t|\theta))) \tag{4}$$

where y_i' is the clustering based hard label and $P_a(x_i^t|\theta)$ and $P_m(x_i^t|\theta)$ are student predictions from both asymmetric branches.

Cross-branch teacher-based soft label training. Clustering algorithms generate hard pseudo labels whose confidences are 100%. Since Re-ID is a fine-grained recognition problem, people with similar clothes are not rare in the dataset. Hard pseudo labels of these similar samples can be extremely noisy. In this case, soft pseudo labels (confidences $< 100\%$) are more reliable. Learning with both hard and soft pseudo labels can effectively alleviate label noise.

The Mean Teacher Model [24] (teacher weights θ') uses the EMA weights of the student model (student weights θ), which shows strong capacity to handle label noise and avoids error amplification along with training. We define θ_t' at training step t as the EMA of successive weights:

$$\theta_t' = \begin{cases} \theta_t, & \text{if } t = 0 \\ \alpha\theta_{t-1}' + (1-\alpha)\theta_t, & \text{otherwise} \end{cases} \tag{5}$$

where α is a smoothing coefficient that controls the self-ensembling speed of the Mean Teacher.

Despite these advantages of Mean Teacher, such self-ensembling teacher-student networks (the teacher is formed by EMA weights of the student, and the student is supervised by the teacher) face the coupling problem. We use the Mean Teacher soft label generator as in [9] and address the coupling problem by cross-branch supervision. Each branch in the student is supervised by a teacher branch which has different structure. Weight diversity between the paired teacher-student can be better kept. Given one target domain sample x_i^t, the teacher (teacher weights θ') encodes it into two feature vectors from two asymmetric branches, average features $F_a(x_i^t|\theta')$ and max features $F_m(x_i^t|\theta')$. The dynamic classifiers then transform these two feature vectors into two predictions respectively

$P_a(x_i^t|\theta')$ and $P_m(x_i^t|\theta')$. Similarly, features of the student (student weights θ) are $F_a(x_i^t|\theta)$ and $F_m(x_i^t|\theta)$, while predictions are $P_a(x_i^t|\theta)$ and $P_m(x_i^t|\theta)$. The predictions from the teacher supervise those from the student with a soft cross-entropy loss [13] in a cross-branch manner, which can be formulated as

$$L_{sce}^{a\to m} = -\sum_i (P_a(x_i^t|\theta') \log(P_m(x_i^t|\theta))) \tag{6}$$

$$L_{sce}^{m\to a} = -\sum_i (P_m(x_i^t|\theta') \log(P_a(x_i^t|\theta))) \tag{7}$$

To further enhance the teacher-student networks' discriminative capacity, the features in the teacher supervise those of the student with a soft triplet loss [9]:

$$L_{stri}^{a\to m} = -\sum_i (T_a(x_i^t|\theta') \log(T_m(x_i^t|\theta))) \tag{8}$$

$$L_{stri}^{m\to a} = -\sum_i (T_m(x_i^t|\theta') \log(T_a(x_i^t|\theta))) \tag{9}$$

where $T(x_i^t|\theta) = \frac{exp(\left\|F(x_i^t|\theta)-F(x_p^t|\theta)\right\|_2)}{exp(\left\|F(x_i^t|\theta)-F(x_p^t|\theta)\right\|_2)+exp(\left\|F(x_i^t|\theta)-F(x_n^t|\theta)\right\|_2)}$ is the softmax triplet distance of the sample x_i^t, its hardest positive x_p^t and its hardest negative x_n^t in a mini-batch. By minimizing the soft triplet loss, the softmax triplet distance in a mini-batch from the student is encouraged to get as close as possible to the distance from the teacher. The positive and negative samples within a mini-batch are decided by clustering-based hard pseudo labels. It can effectively improve the UDA Re-ID performance. The teacher-student networks are trained end-to-end with Equation (4), (6), (7), (8), (9).

$$\begin{aligned} L_{target} = &\lambda_{ce}^t L_{ce} + \lambda_{sce}^t (L_{sce}^{a\to m} + L_{sce}^{m\to a}) \\ &+ \lambda_{stri}^t (L_{stri}^{a\to m} + L_{stri}^{m\to a}) \end{aligned} \tag{10}$$

4. Coupling Problem in Mean Teacher Based Methods

The Mean Teacher Baseline is illustrated in Figure 3 (a) where the student gets supervision from its own EMA weights. In the Mean Teacher Baseline, the student and the teacher quickly converge to each other (coupling problem), which prevents them from exploring more diversified information. Authors of MMT [9] propose to pre-train 2 student networks with different seeds. As illustrated in Figure 3 (b), two Mean Teacher networks are formed separately from two students, which alleviates the coupling problem. However, different initializations decouple two teacher peers only at first epochs. Without a diversity encouragement during the adaptation, two teachers still converge to each other along with training. In Figure 3 (c), our proposed asymmetric branches provide a diversity encouragement during the adaptation, which decouples both teacher peers at all epochs.

To validate our idea, we propose to measure Euclidean distance of appearance signature features between two teacher networks or two teacher branches. We extract feature vectors after global pooling on all images in the target training set. Then, we calculate the Euclidean distance between feature vectors of both teachers and sum up the distance of every image as the final feature distance. If the feature distance is large, we can say that both teacher peers extract diversified features. Otherwise, the teacher peers converge to each other. As we can see from the left curves in Figure 4, the feature distance between two teachers in MMT is large at the beginning, but it decreases and then stabilizes. Differently, the feature distance between two branches in our proposed method is always large during the training. Moreover, we visualize the Euclidean distance of appearance signature features on all target training samples between teacher and student networks in Figure 4 right curves. Our method can maintain a larger distance, which shows that it can better decouple teacher-student networks.

5. Experiments

5.1. Datasets and Evaluation Protocols

Our proposed adaptation method is evaluated on 3 Re-ID datasets: Market→ Duke, Duke → Market, Market → MSMT and Duke → MSMT. **Market-1501** [31] dataset is collected in front of a supermarket in Tsinghua University from 6 cameras. It contains 19,732 images of 751 identities in the training set and 12,936 images of 750 identities in the testing set. **DukeMTMC-reID** [22] is a subset of the DukeMTMC dataset. It contains 16,522 images of 702 persons in the training set, 2,228 query images and 17,661 gallery images of 702 persons for testing from 8 cameras. **MSMT17** [26] is a large-scale Re-ID dataset, which contains 32,621 training images of 1,041 identities and 93,820 testing images of 3,060 identities collected from 15 cameras. Both Cumulative Matching Characteristics (CMC) and mean Average Precisions (mAP) are used in our experiments.

5.2. Implementation details

Hyper-parameters used in our proposed method are searched empirically from the Market→ Duke task and kept the same for the other tasks. To conduct fair comparison with state-of-the-arts, we use a ImageNet [5] pre-trained ResNet-50 [11] as our backbone network. The backbone can be extended to ResNet-based networks designed for cross domain tasks, *e.g.*, IBN-ResNet-50 [19]. An Adam optimizer with a weight decay rate of 0.0005 is used to optimize our networks. Our networks are trained on 4 Nvidia 1080Ti GPUs under Pytorch [20] framework. Detailed con-

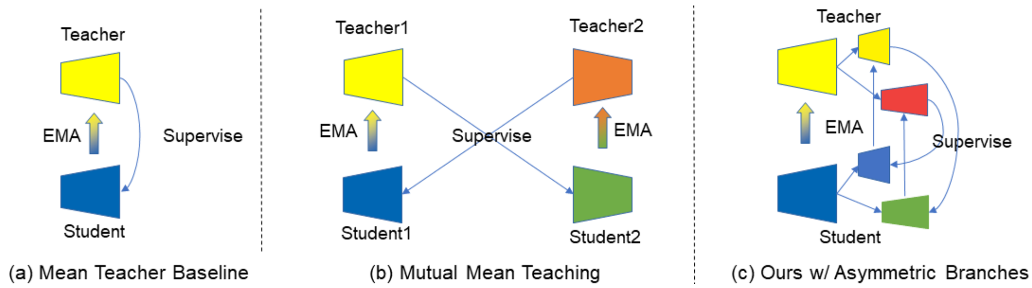

Figure 3. Comparison between (a) Mean Teacher Baseline (b) Mutual Mean Teaching [9] and (c) our Mean Teacher with cross-branch supervised asymmetric branches. Teacher network is formed by exponential moving average (EMA) values of student network.

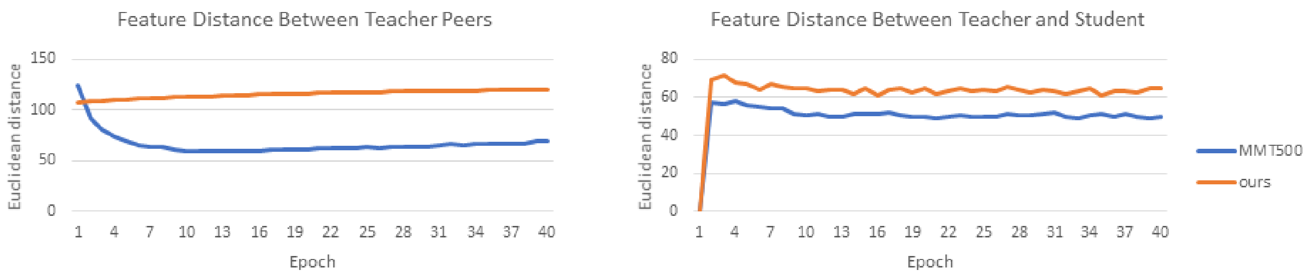

Figure 4. Distance comparison between features extracted from a ResNet50 backbone on all samples in DukeMTMC-reid training set for Market → Duke task. **Left**: Feature distance between two teacher models in MMT and between two teacher branches in our proposed method. **Right**: Feature distance between teacher and student networks.

figurations are given in the following paragraphs.

Stage1: Source domain supervised pre-training. We set $\lambda^s_{ce} = 0.5$ and $\lambda^s_{tri} = 0.5$ in Equation 3. The max epoch E_{pre} is set to 80. For each epoch, the networks are trained $R_{pre} = 200$ iterations.The initial learning rate is set to 0.00035 and is multiplied by 0.1 at the 40th and 70th epoch. For each iteration, 64 images of 16 identities are resized to 256*128 and fed into networks.

Stage2: Target domain unsupervised adaptation. For the clustering, we set the minimum cluster samples to 4 and the density radius p=0.002. Re-ranking parameters for calculating distances are kept the same as in [23] for UDA setting. Re-ranking between source and target domain is not considered for fully unsupervised setting. The Mean Teacher network is initialized and updated in the way of Equation 5 with a smoothing coefficient $\alpha = 0.999$. We set $\lambda^t_{ce} = 0.5$, $\lambda^t_{sce} = 0.5$ and $\lambda^t_{stri} = 1$ in Equation 10. The adaptation epoch E_{ada} is set to 40. For each epoch, the networks are trained $R_{ada} = 400$ iterations with a fixed learning rate 0.00035. For each iteration, 64 images of 16 clustering-based pseudo identities are resized to 256*128 and fed into networks with Random erasing [33] data augmentation.

5.3. Comparison with State-of-the-Art Methods

We compare our proposed methods with state-of-the-art UDA methods in Table 1 for 4 cross-dataset Re-ID tasks: Market→ Duke, Duke → Market, Market → MSMT and Duke → MSMT. Post-processing techniques (*e.g.*, Re-ranking [32]) are not used in the comparison. Our proposed method outperforms MMT [9] (cluster number is set to 500, 700 and 1500 respectively). We can also adjust the density radius in DBSCAN depending on target domain size to get a better performance, but we think it is hard to know the target domain size in the real world. With an IBN-ResNet50 [19] backbone, the performance on 4 tasks can be further improved. Examples of retrieved images are illustrated in Figure 5. Compared to MMT, embeddings from our proposed method contains more discriminative appearance information (*e.g.*, shoulder bag in the first row), which are robust to noisy information (*e.g.*, pose variation in the second row, occlusion in the third row and background variation in the fourth row). This qualitative comparison confirms that appearance signatures of our proposed method are of good quality.

We compare unsupervised Re-ID methods in Table 2. Since the Mean Teacher is designed for handling label noise, it is interesting to see the performance without source pre-training, which introduces more label noise during the adaptation. This setting corresponds to an unsupervised Re-

UDA Methods	Market → Duke		Duke → Market		Market → MSMT		Duke → MSMT	
	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1
HHL (ECCV'18)[34]	27.2	46.9	31.4	62.2	-	-	-	-
UDAP (Arvix'18)[23]	49.0	68.4	53.7	75.8	-	-	-	-
ENC (CVPR'19)[35]	40.4	63.3	43.0	75.1	8.5	25.3	10.2	30.2
PCB-PAST (ICCV'19)[30]	54.3	72.4	54.6	78.4	-	-	-	-
SSG (ICCV'19)[7]	53.4	73.0	58.3	80.0	13.2	31.6	13.3	32.2
ACT (AAAI'20)[28]	54.5	72.4	60.6	80.5	-	-	-	-
MMT500 (ICLR'20)(ResNet50)[9]	63.1	76.8	71.2	87.7	16.6	37.5	17.9	41.3
MMT700 (ICLR'20)(ResNet50)[9]	65.1	78.0	69.0	86.8	-	-	-	-
MMT1500 (ICLR'20)(ResNet50)[9]	-	-	-	-	22.9	49.2	23.3	50.1
ours (ResNet50)	**69.1**	**82.0**	**78.3**	**92.5**	**23.2**	**49.2**	**26.5**	**54.3**
MMT500 (ICLR'20)(IBN-ResNet50)[9]	65.7	79.3	76.5	90.9	19.6	43.3	23.3	50.0
MMT700 (ICLR'20)(IBN-ResNet50)[9]	68.7	81.8	74.5	91.1	-	-	-	-
MMT1500 (ICLR'20)(IBN-ResNet50)[9]	-	-	-	-	26.6	54.4	29.3	58.2
ours (IBN-ResNet50)	**70.8**	**83.3**	**80.4**	**93.0**	**27.8**	**55.5**	**33.0**	**61.8**

Table 1. Comparison of unsupervised domain adaptation (UDA) Re-ID methods (%) on medium-to-medium datasets (Market→ Duke and Duke → Market) and medium-to-large datasets (Market → MSMT and Duke → MSMT).

Unsupervised methods	Market		Duke	
	mAP	Rank1	mAP	Rank1
MMT500*(ICLR'20)[9]	26.9	48.0	7.3	12.7
BUC (AAAI'19)[17]	30.6	61.0	21.9	40.2
SoftSim (CVPR'20)[18]	37.8	71.7	28.6	52.5
TSSL (AAAI'20)[27]	43.3	71.2	38.5	62.2
MMT*+DBSCAN (ICLR'20)[9]	53.5	73.1	54.5	69.5
ours w/o Source pre-training	**65.1**	**82.6**	**63.1**	**77.7**

Table 2. Comparison of unsupervised Re-ID methods (%) with a ResNet50 backbone on Market and Duke datasets. * refers to our implementation where we remove the source pre-training step. DBSCAN refers to a DBSCAN clustering based on re-ranked distance.

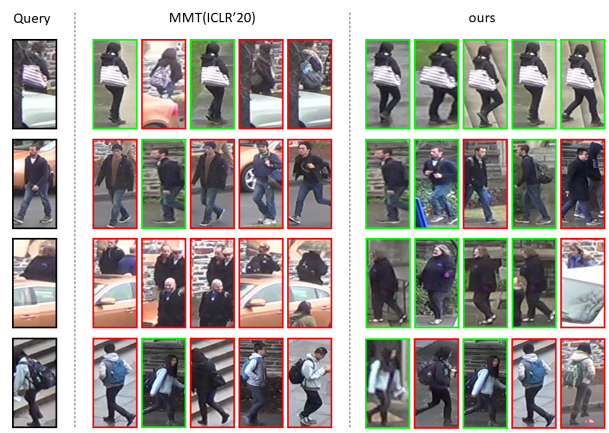

Figure 5. Examples of retrieved most similar 5 images in Market → Duke task from MMT [9] and our proposed method. Given a query image, different identity images are highlighted by red bounding boxes, while same identity images are highlighted by green bounding boxes.

ID. We use ImageNet initialization at the beginning of the adaptation. Our proposed method outperforms previous unsupervised Re-ID by a large margin, which shows that ImageNet initialization can provide basic discriminative capacity for Re-ID.

MMT [9] is the first UDA Re-ID method that uses a Mean Teacher based soft label generator. Authors of MMT propose to use 2 students and 2 teachers with different initialization and stochastic data augmentation to address the coupling problem. We also use Mean Teacher soft pseudo labels but propose a different decoupling solution. Features in asymmetric branches are always extracted in different manners during the adaptation. Compared to MMT, our proposed method has less parameters but achieves better performance. Moreover, in the unsupervised scenario, we can not pre-train MMT with different seeds to obtain different Re-ID initializations. This decoupling strategy becomes inappropriate. Our decoupling strategy relies on structural asymmetry instead of different initializations, which is much more effective in the unsupervised scenario.

ACT [28] uses 2 networks, in which each network learns from its peer. Input data are split into inliers and ouliers after DBSCAN. Then, the first network selects small entropy inliers to train the second network, while the second selects small entropy outliers to train the first. This method enhances input asymmetry by data split. Differently, our proposed method focuses on neural network structure asymmetry. Features are extracted in different ways from same inputs by asymmetric branches, which effectively enhances feature diversity.

5.4. Ablation Studies

Effectiveness of each component in ABMT. Compared with traditional clustering-based Re-ID methods, the performance improvement mainly comes from DBSCAN on re-ranked distance, asymmetric branches and cross-branch supervision. We use a Mean Teacher Baseline where original

Source pre-training	Market → Duke		Duke → Market	
	mAP	Rank1	mAP	Rank1
ResNet50	29.6	46.0	31.8	61.9
ResNet50+AB	31.5	49.7	33.2	63.2
Target adaptation	**Market → Duke**		**Duke → Market**	
	mAP	Rank1	mAP	Rank1
MT-Baseline+K-Means	59.9	74.8	68.9	88.2
MT-Baseline+DBSCAN	61.9	77.3	69.9	88.3
MT-Baseline+K-Means+AB	64.7	78.1	74.8	90.5
MT-Baseline+K-Means+AB+Cross-branch	66.4	79.9	76.8	91.7
MT-Baseline+DBSCAN+AB	67.8	81.1	77.3	92.0
ABMT(MT-Baseline+DBSCAN+AB+Cross-branch)	**69.1**	**82.0**	**78.3**	**92.5**
ABMT+Stochastic data augmentation	68.8	81.2	77.6	91.7
ABMT+Drop out	68.3	81.8	77.9	92.0
ABMT+One more branch	68.1	80.7	76.2	90.4

Table 3. Ablation studies with ResNet50 backbone. MT-Baseline corresponds to the Mean Teacher Baseline in Figure 3 (a) with a ResNet-50. K-Means refers to a K-Means++ clustering whose cluster number is set to 500. AB refers to asymmetric branches. DBSCAN refers to a DBSCAN clustering [6].

ResNet-50 and a K-Means++ clustering of 500 clusters are adopted. We conduct ablation studies by gradually adding one component at each time. Results are shown in Table 3. We can observe: (1) Our proposed asymmetric branches bring the most significant performance improvement during the adaptation. Moreover, as we can see from first two rows in Table 3, they can directly improve the domain generalizability of appearance signatures without target adaptation. (2) DBSCAN on re-ranked distance works better than a K-Means++ clustering of 500 clusters during the adaptation. (3) Cross-branch supervision works on asymmetric branches, which can further improve the adaptation performance.

Can traditional decoupling methods further improve the performance? Enhancing prediction consistency between the teacher and the student under some random noise can effectively improve the performance of SSL. Stochastic data augmentation (teacher inputs and student inputs are under stochastic data augmentation methods) and drop out (teacher feature vectors and student feature vectors are under independent drop out operations before classifiers) are 2 widely-used methods to provide random noise, which also helps to decouple the weights between the teacher and the student. We conduct experiments with stochastic data augmentation (random cropping, random flipping and random erasing) and independent drop out (probability=0.5). The results in Table 3 show that they can not further improve the UDA Re-ID performance. These methods are not designed for fine-grained Re-ID task. When UDA Re-ID performance is already very high, they can not contribute anymore.

Can more branches further improve the performance? We add one more branch to our proposed ABMT. To keep the structural asymmetry in the new branch, the new branch is composed of 5 bottleneck blocks and global average pooling (GAP). We adapt the cross-branch supervision to three branches ($1 \rightarrow 2$, $2 \rightarrow 3$ and $3 \rightarrow 1$). Results are reported in Table 3. The third branch worsens the performance. We argue that the new branch features are not enough distinctive to those from original two branches, which increases the feature duplicateness and worsens the appearance signature quality.

6. Conclusion

In this paper, we propose a novel unsupervised cross-domain Re-ID framework. Our proposed method is mainly based on learning from noisy pseudo labels generated by clustering and Mean Teacher. A self-ensembled Mean Teacher is robust to label noise, but the coupling problem inside paired teacher-student networks leads to a performance bottleneck. To address this problem, we propose asymmetric branches and cross-branch supervision, which can effectively enhance the diversity in two aspects: appearance signature features and teacher-student weights. By enhancing the diversity in the teacher-student networks, our proposed method achieves good performance on both unsupervised domain adaptation and fully unsupervised Re-ID tasks. In future work, we are interested in investigating the performance of other Semi-Supervised Learning methods in unsupervised Re-ID. We are also in exploring the effectiveness of our proposed method in other applications, *e.g.*, Face Recognition.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *SODA '07*, 2007. 1

[2] Hao Chen, Benoit Lagadec, and Francois Bremond. Learning discriminative and generalizable representations by

spatial-channel partition for person re-identification. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 3

[3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 232–242, 2019. 2

[4] Zuozhuo Dai, Mingqiang Chen, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3690–3700, 2018. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 1, 8

[7] Yang Fu, Yunchao Wei, Guanshuo Wang, Xi Zhou, Honghui Shi, and Thomas S. Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6111–6120, 2018. 1, 2, 7

[8] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas S. Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2018. 3

[9] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020. 1, 2, 4, 5, 6, 7

[10] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Wai-Hung Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 1, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 3, 5

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[13] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. 5

[14] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson W. H. Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6727–6735, 2019. 2

[15] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ArXiv*, abs/1610.02242, 2016. 2

[16] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *BMVC*, 2018. 2

[17] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019. 2, 7

[18] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. *ArXiv*, abs/2004.03547, 2020. 1, 2, 7

[19] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 5, 6

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[21] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018. 2

[22] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 5

[23] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *arXiv preprint arXiv:1807.11334*, 2018. 2, 4, 6, 7

[24] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017. 1, 2, 4

[25] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018. 2

[26] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2, 5

[27] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI 2020*, 2020. 2, 7

[28] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross domain person re-identification. 2020. 1, 2, 4, 7

[29] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Wai-Hung Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, 2019. 1

[30] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8221–8230, 2019. 1, 2, 7

[31] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 5

[32] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 2, 4, 6

[33] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *ArXiv*, abs/1708.04896, 2017. 6

[34] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 7

[35] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 7