The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

OVL: One-View Learning for Human Retrieval

Wenjing Li,[1,2] **Zhongcheng Wu**[1,2*]
[1]High Magnetic Field Laboratory, Chinese Academy of Sciences
[2]University of Science and Technology of China

Abstract

This paper considers a novel problem, named One-View Learning (OVL), in human retrieval *a.k.a.* person re-identification (re-ID). Unlike fully-supervised learning, OVL only requires pretty cheap annotation cost: labeled training images are only provided from one camera view (source view/domain), while the annotations of training images from other camera views (target views/domains) are not available. OVL is a problem of multi-target open set domain adaptation that is difficult for existing domain adaptation methods to handle. This is because 1) unlabeled samples are drawn from multiple target views in different distributions, and 2) the target views may contain samples of "unknown identity" that are not shared by the source view. To address this problem, this work introduces a novel one-view learning framework for person re-ID. This is achieved by adversarial multi-view learning (AMVL) and adversarial unknown rejection learning (AURL). The former learns a multi-view discriminator by adversarial learning to align the feature distributions between all views. The later is designed to reject unknown samples from target views through adversarial learning with two unknown identity classifiers. Extensive experiments on three large-scale datasets demonstrate the advantage of the proposed method over state-of-the-art domain adaptation and semi-supervised methods.

1 Introduction

Person re-identification (re-ID) aims to look for the matched person images of the database when given an interested query person. The modern re-ID methods (Li, Zhu, and Gong 2018b; Sun et al. 2018) have achieved impressive improvement in accuracy, relying on rich-labeled data. However, it is time-consuming and difficult to label the identities of persons across disjoint camera views, especially in scenes with a large number of cameras. To mitigate the heavy cost of annotation, many methods for unsupervised domain adaptation (Deng et al. 2018; Wang et al. 2018) are proposed recently. These methods aim at transferring knowledge from a labeled source domain to an unlabeled target domain. Despite their success, these methods still require a large number of labeled auxiliary samples and the utilization of knowledge from the target domain is limited.

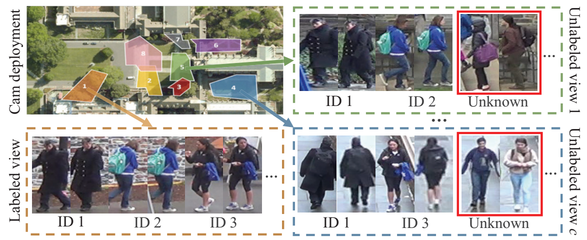


Figure 1: Examples of one-view learning (OVL). Labeled samples are only available from one camera view (source view), while samples of other camera views (target views) are unlabeled. Besides samples of known identities shared by the source view, the target views may contain samples of unknown identity that are absent from the source view.

In the actual labeling process of person re-ID, the main difficulty is matching persons across disjoint camera views. By contrast, it is more easily to label persons under one camera view. This is because: 1) labeling process can be profited from automatic person detection and tracking in raw video of the same camera, and 2) we could avoid the huge effort of finding samples of the same identity across camera views. In light of these advantages, this work considers a novel setting, called one-view learning (OVL), to make trade-off between labeling cost and accuracy for person re-ID. OVL is first introduced by Zhong *et al.* (Zhong et al. 2019), where labeled training samples from one camera view and unlabeled training samples from other camera views are available (Fig. 1). The goal of VOL is to learn a discriminative model that can perform well on testing samples from all views.

OVL can be regarded as a problem of multi-target open set domain adaptation. It has two unique properties that are different from traditional domain adaptation: 1) The unlabeled samples are obtained from multiple unlabeled views (target views/domains) with different distributions. 2) The target views may include samples of identities that are not shared by the labeled view (source view/domain). We call such identities as "*unknown identity*". These two properties

[*]Corresponding author (zcwu@iim.ac.cn).

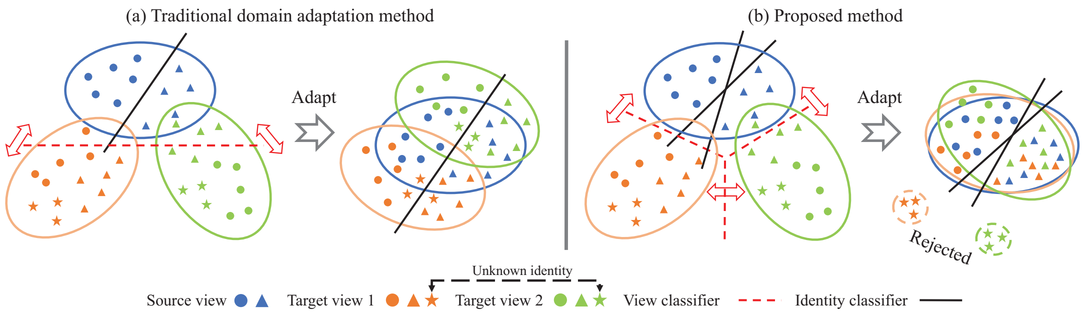


Figure 2: Comparison of traditional domain adaptation method and the proposed method in one-view learning. **(a):** Traditional domain adaptation method mainly attempts to directly align the feature distributions between the source view and the global target view. However, this method may encounter two problems: 1) the gap between target views would still remain, and 2) the samples of unknown identity will be aligned with the source view. **(b):** Our method tries to jointly reduce the gap between all views and reject samples of unknown identity from the target views. Best viewed in color.

make most existing domain adaptation methods (Bousmalis et al. 2017; Ganin and Lempitsky 2015; Tzeng et al. 2017) difficult to solve the problem of OVL. *First*, most works focus on the context of single-source-single-target-domain. Yet unlabeled samples belong to multiple target views in OVL. If we regard target views as one global target view and only focus on reducing the feature distributions between the source view and the global target view, the model may suffer from the variations caused by different target views in testing (Fig. 2(a)). *Second*, most works assume that the source and target views share exactly the same identities/classes. In OVL, however, there may contain samples of unknown identity in the target views. These unknown samples should not be aligned with the source view (Fig. 2(a)). In addition, we do not have any prior knowledge to distinguish unknown samples from the target views. Thus, it is difficult to recognize and reject unknown samples during domain adaptation.

To solve the above difficulties, this work proposes a novel framework for OVL in person re-ID. With respect to the first difficulty, we propose adversarial multi-view learning (AMVL) to align the feature distributions between all views (Fig. 2(b)). AMVL utilizes a multi-view classifier to correctly predict the camera view labels of input samples while encouraging the feature generator to cheat the classifier. This allows the generator to produce view-invariant features for overcoming the variations caused by different views. With respect to the second difficulty, we introduce adversarial unknown rejection learning (AURL) to detect and reject unknown identity samples from the target views (Fig. 2(b)). AURL exploits two unknown identity classifiers to build a decision boundary of unknown identity by enforcing the target samples near to the boundary. On the contrary, the generator attempts to cheat the unknown identity classifiers and push target samples far from the boundary. The generator would choose to 1) align the target samples with the source view or 2) reject them as unknown identity, depending on the output of the unknown identity classifiers.

In summarize, this work makes three contributions. 1) We comprehensively analyze the properties and difficulties of one-view learning (OVL). This helps us to better understand and solve this problem. Moreover, to our knowledge, we are the first to introduce multi-target open set domain adaptation, which is an important problem in real-world applications. 2) We propose a novel and effective method to overcome the difficulties in OVL. Our method jointly considers the divergences between all views and the samples of unknown identity in the target views. Experiment demonstrates the proposed AMVL and AURL are indispensable towards an effective OVL system. 3) Experiment conducted on three large-scale person re-ID datasets shows that our approach achieves state of the art compared with recent unsupervised domain adaptation and semi-supervised methods.

2 Related Work

Unsupervised Domain Adaptation. Unsupervised domain adaptation is mainly divided into two categories: *closed set* domain adaptation and *open set* domain adaptation. Most existing methods focus on the closed set domain adaptation where the source and target domains share exactly the same classes. These methods mainly attempt to align the feature distributions between the source and target domains. For example, reducing the Maximum Mean Discrepancy (MMD) (Gretton et al. 2007) between domains, or, learning an adversarial domain classifier (Ganin and Lempitsky 2015; Tzeng et al. 2017) to produce features that are indistinguishable between the source and target domains. In open set domain adaptation (Busto and Gall 2017), there may exist samples of unknown class in the target domain. In this situation, the traditional distribution matching approaches may not be suitable. Because the samples of unknown class should not be aligned with the source domain. To address this problem, recent methods (Baktashmotlagh et al. 2019; Busto and Gall 2017; Saito et al. 2018) aim to detect and reject unknown class samples during distribution alignment. Saito *et al.* (Saito et al. 2018) employ adversarial training to build an unknown class decision boundary and separate the unknown target samples from known ones. Baktashmotlagh *et al.* (Baktashmotlagh et al. 2019) propose a framework that disentangles the data into shared and private representations. The unknown class samples are detected through estimating whether the data can be reconstructed by the private representation. Although many works have been proposed for multi-source domain adaptation (Mansour, Mohri, and Rostamizadeh 2009; Li, Carlson, and others 2018; Zhao et al. 2018), there is only one work studies on multi-target domain adaptation (Gholami et al. 2018). In this paper, we consider a more challenging setting, multi-target open set domain adaptation, where we not only need to align the distributions between all domains, but also detect and reject samples of unknown class from the target domains.

Person re-identification. Recent methods have made great achievement in fully-supervised person re-identification (re-ID) (Li, Zhu, and Gong 2018b; Sun et al. 2018), benefiting from the rich-annotated data. However, labeling person re-ID data across disjoint cameras is a time-consuming and labor demanding process. To overcome this problem, recent works focus on studies of unsupervised learning (Chen, Zhu, and Gong 2018; Yang et al. 2017; 2014), semi-supervised learning (Li, Zhu, and Gong 2018a; Liu, Wang, and Lu 2017; Ye et al. 2017) and unsupervised domain adaptation (Fan et al. 2018; Wang et al. 2018; Zhong et al. 2018). Although SMP (Liu, Wang, and Lu 2017) and DGM (Ye et al. 2017) claim that they are unsupervised methods, they are in fact semi-supervised methods and only can be implemented in video-based person re-ID. Since they need to assign at least one tracklet for each identity. TAUDL (Li, Zhu, and Gong 2018a) proposes an unsupervised method for video-based person re-ID. However, it is a semi-supervised method for image-based person re-ID. Because TAUDL assigns all person images per ID per camera to a unique label. OVL can also be termed as a semi-supervised problem where samples of unknown identity may exist in the unlabeled samples. However, this work attempts to solve OVL in view of domain adaptation. For unsupervised domain adaptation in person re-ID, the identities from the source and target domains are completely different. Thus, it is improper to directly align the distributions of domains in the space of identity. To solve this problem, recent methods mainly try to align the source and target domains in the pixel-level space (Deng et al. 2018; Wei et al. 2018) or attribute-level space (Lin et al. 2018; Wang et al. 2018). Compared to the above unsupervised domain adaptation methods, in OVL, the target domains may contain samples of known/unknown identities that are shared/unshared by the labeled source domain. Therefore, we can address the problem of OVL by aligning feature distributions of domains in the identity space, but should notice and reject the unknown identity samples from known ones.

3 Method

3.1 Problem Definition of One-View Learning

In one-view learning (OVL), we are provided with a training dataset collected from C camera views. The training data includes labeled and unlabeled samples. The labeled data is only collected from one camera view whereas the unlabeled data is captured from other $C-1$ camera views. We regard labeled training data $\{X_s, Y_s\}$ as source view/domain, which includes N_s person images. The number of identities in the source view is M. We define these identities as known identities. Since the unlabeled data is drawn from $C-1$ camera views, we divide it into $C-1$ target views/domains. For each target view $X_{t,c}$ belonging to camera c, we are provided with $N_{t,c}$ unlabeled person images. The goal of OVL is to learn a model using samples of the source and target views, so that the model could extract discriminative representation on the testing set. In testing, person images are draw from all C camera views.

OVL is a problem of multi-target open set domain adaptation which has the following two properties: 1) Training samples are draw from one labeled source view and $C-1$ unlabeled target views. 2) A person would not always appear under all cameras. Therefore, there may contain samples of unknown identity in the target views. The unknown identity indicates the persons that are absent from the source view. Based on these two properties, this work aims to address two difficulties that are hard for traditional domain adaptation methods. *First*, instead of directly reducing the distribution gap between the source view and the global target view, we should also consider the distribution gap between each pair of target views. This is because we need to compare the similarities between samples from all C views during testing. *Second*, the target views may contain samples of unknown identity that should not be aligned with the source view. Thus, we need to detect unknown identity samples from the target views and reject them during adapting. Next, we will introduce our approach to address the above difficulties for OVL.

3.2 Overview of The Framework

The framework of our method is shown in Fig 3. The input of the network is the samples of the labeled source view and unlabeled target views. Our network is comprised of four modules: a feature generator (G), two identity classifiers ($F_{I,1}$ and $F_{I,2}$), and a multi-view classifier (F_C). The generator is composed of several residual blocks (He et al. 2016). The module of classifier has two fully convolutional (FC) layers. The output is $M+1$-dimensional for identity classifier and C-dimensional for multi-view classifier. The outputs of classifiers are obtained by softmax activation function. The first M dimensions of the output for identity classifier are the predicted probabilities of known identities while the last dimension represents the predicted probability of unknown identity. We initialize $F_{I,1}$ and $F_{I,2}$ differently to create two different identity classifiers. During training, we introduce three learning strategies to optimize the network, *i.e.* supervised learning, adversarial multi-view learning and adversarial unknown rejection learning. The supervised learning is implemented on the labeled source view. It aims to learn basic discriminative feature generator and identity classifiers using the identity label of the source data. The adversarial multi-view learning (AMVL) is proposed to reduce the gap between all views by training the multi-view classifier with adversarial learning. The adversarial unknown rejection learning (AURL) is introduced to reject unknown identity samples during adapting process. The two identity classifiers attempt to make a decision boundary of unknown

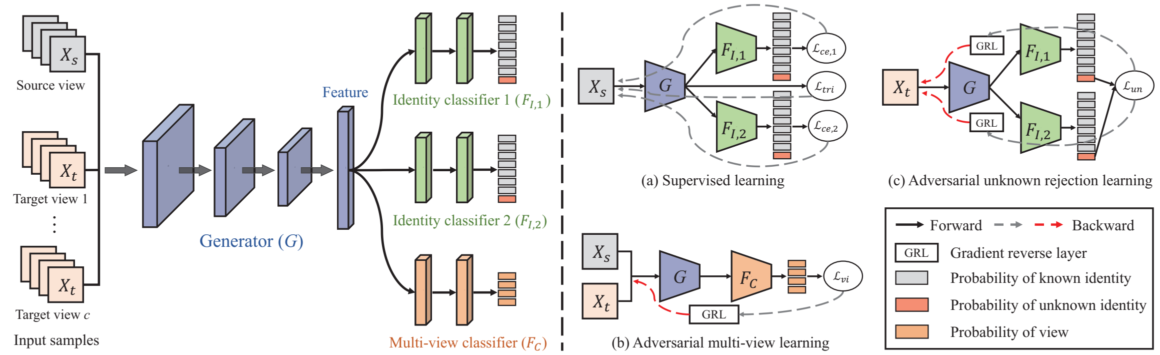


Figure 3: The framework of the proposed method. **Left:** The network of the proposed method. Given the samples of the labeled source view and unlabeled target views, we forward them into the network. The network has four modules: a feature generator (G), two identity classifiers ($F_{I,1}$ and $F_{I,2}$), and a multi-view classifier (F_C). **Right:** The loss and optimization of the proposed method. During training, we jointly perform supervised learning, adversarial multi-view learning and adversarial unknown rejection learning to optimize the network. $\mathcal{L}_{ce}$ and $\mathcal{L}_{tri}$ indicate the cross-entropy loss and triplet loss for labeled source samples, respectively. $\mathcal{L}_{vi}$ represents the view classification loss for source and target samples. $\mathcal{L}_{un}$ denotes the unknown rejection loss of target samples.

identity by enforcing the target samples near to the unknown boundary. By contrast, the generator tries to push the target samples away from the boundary depending on the probability of the unknown identity. Next, we will introduce the optimization of the proposed method in detail.

3.3 Supervised Learning on Source View

Given the labeled source samples, we are able to train the network in a supervised way. As shown in Fig 3(a), we adopt classification loss and triplet loss (Hermans, Beyer, and Leibe 2017) to perform the supervised learning on the source view:

$$\begin{aligned}\mathcal{L}_{sl} = \mathcal{L}_{ce,1}(F_{I,1}(G(x_s))) + \mathcal{L}_{ce,2}(F_{I,2}(G(x_s))) \\ + \mathcal{L}_{tri}(G(x_s)),\end{aligned} \tag{1}$$

where $\mathcal{L}_{ce,1}$ and $\mathcal{L}_{ce,2}$ denote the cross-entropy losses with respect to F_1 and F_2, respectively. $\mathcal{L}_{ce,j}$ is formulated as:

$$\mathcal{L}_{ce,j} = -\log p_j(y_s|x_s), \tag{2}$$

where $p_j(y_s|x_s)$ is the probability of identity label for the input x_s predicted by the classifier $F_{I,j}$. The triplet loss is explained as,

$$\mathcal{L}_{tri} = [m + D(x_s, x_{s,p}) - D(x_s, x_{s,n})], \tag{3}$$

where $x_{s,p}$ and $x_{s,n}$ represent the positive sample and negative sample of the input x_s in the training batch. m is a margin parameter and $D(\cdot)$ is the Euclidean distance between two features obtained by the generator G. We empirically set m to 0.3 in this paper.

3.4 Adversarial Multi-View Learning

Due to the distribution divergences between the source and target views, the network trained on the source view may fail to extract discriminative feature for the target views. As discussed in the first difficulty of Sec. 3.1, it is important to reduce the distribution gap between each pair of views. To achieve this goal, we propose adversarial multi-view learning (AMVL) to align the feature distributions between all views. As shown in Fig 3(b), we propose to utilize the cross-entropy loss on the output of the multi-view classifier F_C:

$$\mathcal{L}_{vi} = -\log q(c|x), \tag{4}$$

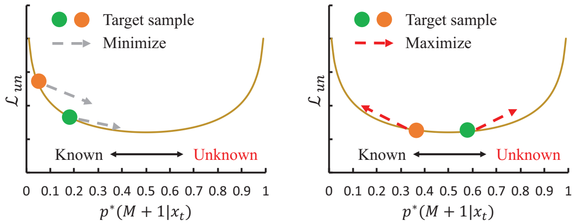
where $q(c|x)$ is the probability of camera view label for input source/target sample x obtained by the multi-view classifier F_C. Then, we apply the adversarial training to optimize the generator and the multi-view classifier. The training object of $\mathcal{L}_{vi}$ is,

$$\max_{G} \min_{F_C} \mathcal{L}_{vi}. \tag{5}$$

The multi-view classifier attempts to correctly predict the camera view label of the input sample whereas the generator tries to cheat the multi-view classifier. In this way, the generator is encouraged to produce the feature that is indistinguishable by the multi-view classifier. Thereby, the feature distributions of all views could be aligned and the generator is able to produce view-invariant features.

3.5 Adversarial Unknown Rejection Learning

As mentioned in the second difficulty of Sec. 3.1, the target views may include samples of unknown identity that are not shared by the source view. The samples of unknown identity should not be aligned with the source view and thus should be rejected during adapting process. Inspired by (Saito et al. 2018), we propose to construct a decision boundary for the unknown identity. The decision boundary of the unknown identity is utilized to detect and reject the samples of unknown identity. We first try to build a decision boundary for unknown identity by enforcing the target samples near to the decision boundary with the identity classifiers. Then, we train the feature generator to cheat the classifiers. The feature generator has two choices, pushing the target samples from the decision boundary to the side of unknown identity



(a) Optimization of identify classifiers

(b) Optimization of generator

Figure 4: Examples of adversarial unknown rejection learning. (a) The identify classifiers try to push the target samples to near the unknown boundary. (b) The generator tries to distinguish unknown target samples from known ones.

or known identity. Specifically, we train to classify the target samples to unknown identity using $F_{I,1}$ and classify the target samples to known identity using $F_{I,2}$. The loss function is formulated as,

$$\mathcal{L}_{un} = -\log p_1(M+1|x_t) - \log(1-p_2(M+1|x_t)), \quad (6)$$

where $p_1(M+1|x_t)$ and $p_2(M+1|x_t)$ represent the $M+1th$ dimension of the output obtained by $F_{I,1}$ and $F_{I,2}$, respectively. We utilize adversarial training to optimize this object,

$$\max_{G} \min_{F_{I,1},F_{I,2}} \mathcal{L}_{un}. \quad (7)$$

Since we use exactly the same source samples to train $F_{I,1}$ and $F_{I,2}$, these two identity classifiers would converge to similar parameters. In this way, the outputs of the $F_{I,1}$ and $F_{I,2}$ would be approximately equal. We replace $p_1(M+1|x_t)$ and $p_2(M+1|x_t)$ by $p^*(M+1|x_t)$ to help us understand the optimization of AURL. The object of Eq. 6 can be reformulated as,

$$\mathcal{L}_{un} = -\log p^*(M+1|x_t) - \log(1-p^*(M+1|x_t)). \quad (8)$$

As shown in Fig. 4, the minimization of $\mathcal{L}_{un}$ is $p^*(M+1|x_t) = 0.5$. Therefore, the classifiers try to push the value of $p^*(M+1|x_t)$ to 0.5. On the contrary, the generator tries to maximize $\mathcal{L}_{un}$ and thus encouraging the value of $p^*(M+1|x_t)$ far from 0.5. In this way, the generator has two options: pushing the target sample as unknown identity if $p^*(M+1|x_t)$ is larger than 0.5, and vice versa.

3.6 Overall Optimization

Taking into account the supervised learning, adversarial multi-view learning and adversarial unknown rejection learning, the overall objectives of the proposed method are:

$$\begin{gathered}\min_{G,F_{I,1},F_{I,2}} \mathcal{L}_{sl}, \\ \max_{G} \min_{F_{I,1},F_{I,2},F_C} \lambda_{vi}\mathcal{L}_{vi} + \lambda_{un}\mathcal{L}_{un},\end{gathered} \quad (9)$$

where λ_{vi} and λ_{un} are hyper-parameters that control the importance of AMVL and AURL, respectively. We utilize the gradient reverse layer (Ganin and Lempitsky 2015) to efficiently implement the adversarial training in one step.

4 Experiment

4.1 Dataset and Implement Details

Dataset. We evaluate the proposed method on three large-scale person re-ID benchmarks: Market-1501 (Zheng et al. 2015), DukeMTMC-reID (Ristani et al. 2016; Zheng, Zheng, and Yang 2017) and MSMT17 (Wei et al. 2018). Performance is evaluated by the cumulative matching characteristic (CMC) and mean Average Precision (mAP).

Network. In this paper, we utilize ResNet-50 (He et al. 2016) (without classifier layers) initialized on the ImageNet (Deng et al. 2009) as the backbone of the generator. The classifier module is composed of two fully convolutional (FC) layers. The first FC layer is 1024-dimensional. The second FC layer is the classification layer which is $M+1$-dimensional for identity classifier and C-dimensional for multi-view classifier. We resize the input image to 256 ×128. The random flipping and random cropping are applied for data augmentation during training. We initialize the learning rate to 0.01 for the generator and 0.1 for the classifiers. The learning rate is divided by 10 after 40 epochs. The batch size is set to 64 for both source and target views. The SGD optimizer is used to train the network in total of 60 epochs. In default, we set $\lambda_{vi} = 0.2$ and $\lambda_{un} = 0.1$. In testing, we extract the L2-normalized output of generator as the image feature. The similarities between query and gallery images are calculated through Euclidean distance. Note that, the testing samples are drawn from all views.

Fully-supervised learning uses fully-labeled data to train the network with the supervised learning loss. Namely, the identities of training samples in all camera views are available. **Baseline** uses only the labeled source view data to train the network with the supervised learning loss.

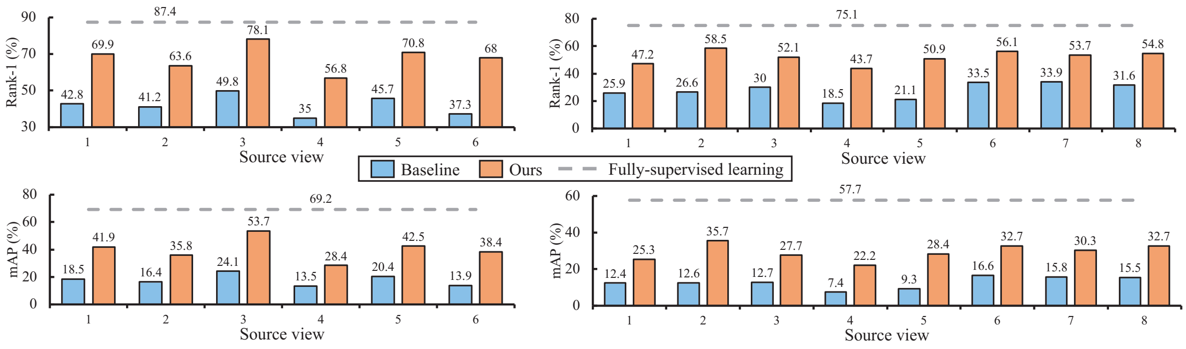
4.2 Parameter Analysis

We first analyze the sensitivities of the weights of adversarial multi-view learning and adversarial unknown rejection learning. We vary the value of one weight and keep another fixed. To avoid over-adjusting the parameters, we only evaluate the weights on one labeled source view. Specifically, we use the $3th$ view and $2th$ view as the source views for Market-1501 and DukeMTMC-reID, respectively.

Weight of adversarial multi-view learning. The evaluation of different values of λ_{vi} is shown in Table 1. When $\lambda_{vi} = 0$, the model is trained without $\mathcal{L}_{vi}$. After injecting adversarial multi-view learning into the system, the performance is consistently improved when λ_{vi} is in range $[0.1, 0.5]$. Assigning a large value to λ_{vi} will decrease the performance. The best results are obtained when $\lambda_{vi} = 0.2$.

Weight of adversarial unknown rejection learning. In Table 2, we evaluate the impact of λ_{un}. When $\lambda_{un} = 0$, our method reduces to the model trained with supervised learning and adversarial multi-view learning. It can be seen that, when adding adversarial unknown rejection learning into the system ($\lambda_{un} > 0$), the rank-1 accuracy and mAP improve with the increase of λ_{un} and achieve best results when λ_{un} is around 0.1.

In the following experiments, we set $\mathcal{L}_{vi} = 0.2$ and $\mathcal{L}_{un} = 0.1$ for all settings.



(a) Market-1501

(b) DukeMTMC-reID

Figure 5: Results of training the model using different source views on Market-1501 and DukeMTMC-reID.

Table 1: Evaluation with different values of λ_{vi} on Market-1501 and DukeMTMC-reID. We fix λ_{un} to 0.1.

λ_{vi}	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
0.0	72.2	47.6	52.0	31.3
0.1	75.1	49.6	58.1	35.6
0.2	**78.1**	**53.7**	**58.5**	**35.7**
0.5	74.4	48.4	57.1	34.9
1.0	73.5	46.9	50.8	29.7

Table 2: Evaluation with different values of λ_{un} on Market-1501 and DukeMTMC-reID. We fix λ_{vi} to 0.2.

λ_{un}	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
0.0	70.3	42.3	53.0	30.9
0.01	72.7	46.3	54.6	31.3
0.05	76.4	52.6	57.3	35.5
0.1	**78.1**	**53.7**	**58.5**	**35.7**
0.5	73.0	47.3	54.3	31.9

Table 3: Ablation study of our approach on Market-1501 and DukeMTMC-reID. **Average**: Average results on all source views. **Max**: The best results over all source views. The best results are achieved by 3*th* view for Market-1501 and 2*th* view for DukeMTMC-reID.

Methods			Market-1501		Duke	
			R-1	mAP	R-1	mAP
Fully-Supervised Learning			87.4	69.2	75.1	57.7
One-View Learning	Average	Baseline	41.9	17.8	27.6	12.7
		Ours w/o $\mathcal{L}_{tri}$	59.5	30.3	43.4	21.5
		Ours w/o $\mathcal{L}_{vi}$	61.7	35.4	46.0	25.6
		Ours w/o $\mathcal{L}_{un}$	62.0	31.9	48.0	25.1
		Ours	**67.8**	**40.1**	**52.1**	**29.3**
	Max	Baseline	49.8	24.1	33.9	15.8
		Ours	**78.1**	**53.7**	**58.5**	**35.7**

Table 4: Comparison of traditional distribution matching methods and AMVL. Results averaged on all source views are reported. V1: Align the source view with the global target view; V2: Adapt the source view to each target view with $C-1$ domain classifiers.

Method	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
Baseline	41.9	17.8	27.6	12.7
Basel.+DANN (V1)	52.4	24.7	32.7	15.8
Basel.+ADDA (V1)	53.4	26.6	33.2	16.5
Basel.+DANN (V2)	55.2	28.1	39.2	19.3
Basel.+ADDA (V2)	55.6	28.8	40.5	21.2
Basel.+AMVL	**62.0**	**31.9**	**48.0**	**25.1**

4.3 Evaluation

We conduct detailed evaluations of our method on Market-1501 and DukeMTMC-reID in Fig. 5 and Table 3.

Performance of the baseline. We first evaluate the results of baseline in OVL. As shown in Fig. 5 and Table 3, the results of baseline are largely lower than that of fully-supervised learning. This is because that the baseline only uses limited labeled data from one view to train the model. Without learning with samples of other views, the model would be significantly suffered from the variations caused by unseen camera views.

Performance of the proposed method. We then evaluate the effectiveness of the proposed method in Fig 5 and Table 3. It is clear that our method consistently improves the results of baseline by a large margin in all settings. Specifically, our approach improves the average rank-1 accuracy of all source views by 25.9% for Market-1501 and 24.5% for DukeMTMC-reID. The best results are achieved when using 3*th* view and 2*th* view as the source views for Market-1501 and for DukeMTMC-reID, respectively. Our method achieves 78.1% in rank-1 accuracy with 2,707 labeled samples of 694 identities when tested on Market-1501. This is 9.3% lower than fully-supervised learning which uses 12,936 labeled samples of 751 identities.

Ablation experiment on the proposed method. We further investigate the importance of the components in our method. First, as shown in Table 3, the triplet loss $\mathcal{L}_{tri}$ in supervised learning is effective to improve the performance

Table 5: Comparison with state-of-the-art domain adaptation methods and semi-supervised methods. *: Domain adaptation methods benefited from extra labelled auxiliary training data. †: Reproduced by this paper with the setting of one-view learning.

Method	Reference	Market-1501		DukeMTMC-ReID		MSMT17	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CAMEL* (Yu, Wu, and Zheng 2017)	ICCV 2017	54.5	26.3	-	-	-	-
PUL* (Fan et al. 2018)	TOMM 2018	44.7	20.1	30.4	16.4	-	-
PTGAN* (Wei et al. 2018)	CVPR 2018	38.6	-	27.4	-	11.8	3.3
SPGAN* (Deng et al. 2018)	CVPR 2018	51.5	22.8	41.1	22.3	-	-
SPGAN+LMP* (Deng et al. 2018)	CVPR 2018	57.7	26.7	46.4	26.2	-	-
TJ-AIDL* (Wang et al. 2018)	CVPR 2018	58.2	26.5	44.3	23.0	-	-
HHL* (Zhong et al. 2018)	ECCV 2018	62.2	31.4	46.9	27.2	-	-
DAS* (Bak, Carr, and Lalonde 2018)	ECCV 2018	65.7	-	-	-	-	-
TAUDL (Li, Zhu, and Gong 2018a)	ECCV 2018	63.7	41.2	61.7	43.5	28.4	12.5
EUG† (Wu et al. 2018)	CVPR 2018	69.8	44.7	37.8	18.7	11.9	3.0
CamStyle (Zhong et al. 2019)	TIP 2019	67.0	38.6	54.9	30.8	-	-
Ours	AAAI 2020	**78.1**	**53.7**	**58.5**	**35.7**	**33.9**	**11.3**

in OVL. For example, when removing triplet loss $\mathcal{L}_{tri}$ from our model, the average rank-1 accuracy drops from 67.8% to 59.5% for Market-1501. A similar phenomenon is observed on DukeMTMC-reID.

Next, we validate the effectiveness of the adversarial multi-view learning (AMVL). As reported in Table 3, AMVL is indispensable to reduce the gap between different views. For example, without AMVL, the results of our method drop by 8.3% for Market-1501 and 8.7% for DukeMTMC-reID in average rank-1 accuracy, respectively. In addition, we compare AMVL with two popular distribution matching methods in domain adaptation, *i.e.* DANN (Ganin and Lempitsky 2015) and ADDA (Tzeng et al. 2017). We implement them in two ways: 1) align the source view with the global target view, and 2) adapt the source view to each target view with $C-1$ domain classifiers. These two ways only focus on reducing the distribution gap between the source view and the target view while ignoring the distribution gap between each pair of target views. As shown in Table 4, AMVL clearly outperforms DANN and ADDA. This demonstrates the importance of aligning the feature distribution between target views.

Finally, we evaluate the effect of adversarial unknown rejection learning (AURL). In Table 3, we observe consistent improvement when adding AURL into the system. For example, when only injecting AURL into the baseline, "Ours w/o $\mathcal{L}_{vi}$" improves the average rank-1 accuracy by 19.8% for Market-1501 and by 18.4% for DukeMTMC-reID. This indicates that AURL helps to align the target views with the source view. Moreover, when given a model trained with AMVL ("Ours w/o $\mathcal{L}_{un}$"), AURL mainly focuses on avoiding aligning target samples of unknown identity with the source view. This helps us to further improve the results of the system. For instance, when tested on Market-1501, the baseline trained with AMVL and AURL ("Ours") achieves 67.8% in average rank-1 accuracy, improving the average rank-1 accuracy of "Ours w/o $\mathcal{L}_{un}$" by 5.8%.

4.4 Comparison with State-of-the-art Methods

In Table 5, we compare with 10 state-of-the-art methods including 7 unsupervised domain adaptation methods (CAMEL (Yu, Wu, and Zheng 2017), PUL (Fan et al. 2018), PTGAN (Wei et al. 2018), SPGAN (Deng et al. 2018), TJ-AIDL (Wang et al. 2018), HHL (Zhong et al. 2018), DAS (Bak, Carr, and Lalonde 2018)) and 3 semi-supervised methods (TAUDL (Li, Zhu, and Gong 2018a), EUG (Wu et al. 2018), CamStyle (Zhong et al. 2019)). Results are evaluated on Market-1501, DukeMTMC-reID and MSMT17. The unsupervised domain adaptation methods aim to transfer the knowledge (identity/attribute) from extra labelled auxiliary training dataset to an unlabeled target dataset. In general, the extra auxiliary dataset and the target dataset are draw from different distributions. The semi-supervised methods aim to leverage limited labeled samples and a large number of unlabeled samples to learn a discriminative model. Although TAUDL claims that it is an unsupervised method, it actually is a semi-supervised method when implemented on image-based datasets. Because TAUDL assigns all person images per ID per camera to a unique label in a camera-independent manner. Instead of using camera-independent labeled samples in all camera views, one-view learning (OVL) only requires labeled samples from one camera view. We reproduce EUG in the setting of OVL. For OVL, we use the 3*th* view, 2*th* view and 1*th* view as the source views for Market-1501, DukeMTMC-reID and MSMT17, respectively.

As shown in Table 5, our approach outperforms all domain adaptation methods by a large margin. For example, our approach surpasses HHL by 15.9% for Market-1501 and by 11.6% for DukeMTMC-reID in rank-1 accuracy. It is worth noting that our approach does not require any extra labelled auxiliary training data as compared to HHL. Instead, our approach only uses limited labeled data of one camera view which can be easily obtained. When using the same training samples (OVL), our approach clearly outperforms CamStyle and EUG on all datasets. The main reason of the inferior of EUG is that EUG gradually predicts pseudo label for unlabeled data but ignores the existing of unknown identity samples. Assigning known identities to unknown identity samples is unreasonable and would undoubtedly harm the performance of the model. For example, when tested on DukeMTMC-reID and MSMT17, EUG fails to produce competitive results. Because the source view includes much less identities than the overall dataset. Compared to CamStyle which requires to learn a lot of complicated style-transferred models, our method is easy to implement and produces higher results than that of CamStyle. Our approach is significantly superior to TAUDL on Market-1501 and achieves competitive results with TAUDL on MSMT17. Although our approach obtains lower results than TAUDL on DukeMTMC-reID, the labeled samples used in our approach are much less than that used in TAUDL.

5 Conclusion

In this paper, we consider a novel setting, one-view learning (OVL), for person re-identification (re-ID). OVL is an important and practical problem in balancing the annotation cost and accuracy for person re-ID. This work comprehensively investigates the properties and difficulties of OVL and proposes an effective framework to address these difficulties. Specifically, we introduce adversarial multi-view learning (AMVL) and adversarial unknown rejection learning (AURL) to reduce the distribution gap between all views and reject unknown identity samples during adapting. Experiments on three datasets demonstrate the effectiveness of the proposed method and show that our approach could achieve state of the art compared with the advanced unsupervised domain adaptation and semi-supervised methods.

References

Bak, S.; Carr, P.; and Lalonde, J.-F. 2018. Domain adaptation through synthesis for unsupervised person re-identification. In *Proc. ECCV*.

Baktashmotlagh, M.; Faraki, M.; Drummond, T.; and Salzmann, M. 2019. Learning factorized representations for open-set domain adaptation. In *Proc. ICLR*.

Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; and Krishnan, D. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proc. CVPR*.

Busto, P. P., and Gall, J. 2017. Open set domain adaptation. In *Proc. ICCV*.

Chen, Y.; Zhu, X.; and Gong, S. 2018. Deep association learning for unsupervised video person re-identification. In *Proc. BMVC*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*.

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 *Proc. CVPR*.

Fan, H.; Zheng, L.; Yan, C.; and Yang, Y. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM TOMM*.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *Proc. ICML*.

Gholami, B.; Sahu, P.; Rudovic, O.; Bousmalis, K.; and Pavlovic, V. 2018. Unsupervised multi-target domain adaptation: An information theoretic approach. *arXiv*.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2007. A kernel method for the two-sample-problem. In *Proc. NeurIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. CVPR*.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv*.

Li, Y.; Carlson, D. E.; et al. 2018. Extracting relationships by multi-domain matching. In *Proc. NeurIPS*.

Li, M.; Zhu, X.; and Gong, S. 2018a. Unsupervised person re-identification by deep learning tracklet association. In *Proc. ECCV*.

Li, W.; Zhu, X.; and Gong, S. 2018b. Harmonious attention network for person re-identification. In *Proc. CVPR*.

Lin, S.; Li, H.; Li, C.-T.; and Kot, A. C. 2018. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *Prco. BMVC*.

Liu, Z.; Wang, D.; and Lu, H. 2017. Stepwise metric promotion for unsupervised video person re-identification. In *Proc. ICCV*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009. Domain adaptation with multiple sources. In *Proc. NeurIPS*.

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 *Proc. ECCVW*.

Saito, K.; Yamamoto, S.; Ushiku, Y.; and Harada, T. 2018. Open set domain adaptation by backpropagation. In *Proc. ECCV*.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proc. ECCV*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proc. CVPR*.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proc. CVPR*.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proc. CVPR*.

Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; and Yang, Y. 2018. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proc. CVPR*.

Yang, Y.; Yang, J.; Yan, J.; Liao, S.; Yi, D.; and Li, S. Z. 2014. Salient color names for person re-identification. In *Proc. ECCV*.

Yang, Y.; Wen, L.; Lyu, S.; and Li, S. Z. 2017. Unsupervised learning of multi-level descriptors for person re-identification. In *Proc. AAAI*.

Ye, M.; Ma, A. J.; Zheng, L.; Li, J.; and Yuen, P. C. 2017. Dynamic label graph matching for unsupervised video re-identification. In *Proc. ICCV*.

Yu, H.-X.; Wu, A.; and Zheng, W.-S. 2017. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proc. ICCV*. IEEE.

Zhao, H.; Zhang, S.; Wu, G.; Moura, J. M.; Costeira, J. P.; and Gordon, G. J. 2018. Adversarial multiple source domain adaptation. In *Proc. NeurIPS*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proc. ICCV*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. ICCV*.

Zhong, Z.; Zheng, L.; Li, S.; and Yang, Y. 2018. Generalizing a person retrieval model hetero- and homogeneously. In *Proc. ECCV*.

Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; and Yang, Y. 2019. Camstyle: A novel data augmentation method for person re-identification. *IEEE TIP*.