CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification

Guillaume Delorme, Yihong Xu, Stéphane Lathuilière, Radu Horaud, Xavier Alameda-Pineda

CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification

Guillaume Delorme[1] Yihong Xu[1] Stéphane Lathuilière[2] Radu Horaud[1] Xavier Alameda-Pineda[1]
[1]Inria, LJK, Univ. Grenoble Alpes, France [2] LTCI, Télécom Paris, IP Paris, France
Email: firstname.lastname@{[1]inria.fr [2]telecom-paris.fr}

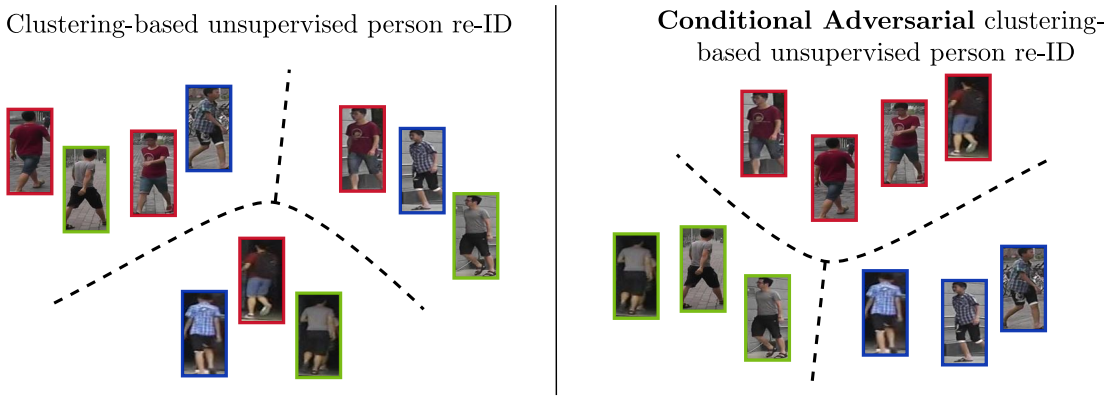


Fig. 1: Clustering-based (left) vs. conditional adversarial clustering-based (right) unsupervised person re-ID. Our intuition is that features should be camera-independent, and thus the clustering result should group visual features by ID rather than by camera. Our method conditions a camera-based adversarial discriminator with the visual features corresponding to the cluster's centroid.

***Abstract*—Unsupervised person re-ID is the task of identifying people on a target data set for which the ID labels are unavailable during training. In this paper, we propose to unify two trends in unsupervised person re-ID: clustering & fine-tuning and adversarial learning. On one side, clustering groups training images into pseudo-ID labels, and uses them to fine-tune the feature extractor. On the other side, adversarial learning is used, inspired by domain adaptation, to match distributions from different domains. Since target data is distributed across different camera viewpoints, we propose to model each camera as an independent domain, and aim to learn domain-independent features. Straightforward adversarial learning yields negative transfer, we thus introduce a conditioning vector to mitigate this undesirable effect. In our framework, the centroid of the cluster to which the visual sample belongs is used as conditioning vector of our conditional adversarial network, where the vector is permutation invariant (clusters ordering does not matter) and its size is independent of the number of clusters. To our knowledge, we are the first to propose the use of conditional adversarial networks for unsupervised person re-ID. We evaluate the proposed architecture on top of two state-of-the-art clustering-based unsupervised person re-identification (re-ID) methods on four different experimental settings with three different data sets and set the new state-of-the-art performance on all four of them. Our code and model will be made publicly available at https://team.inria.fr/perception/canu-reid/.**

I. Introduction

Person re-identification (re-ID) is a well-studied retrieval task [1]–[3] that consists in associating images of the same person across cameras, places and time. Given a query image of a person, we aim to recover his/her identity (ID) from a set of identity-labeled gallery images. The person re-ID task is particularly challenging for two reasons. First, query and gallery images contain only IDs which have never been seen during training. Second, gallery and query images are captured under a variety of background, illumination, viewpoints and occlusions.

Most re-ID models assume the availability of heavily labeled datasets and focus on improving their performance on the very same data sets, see for instance [4], [5]. The limited generalization capabilities of such methods were pointed out in previous literature [6], [7]. In the recent past, researchers attempted to overcome this limitation by investigating a new person re-ID task, where there is a *source* dataset annotated with person IDs and another unlabeled *target* dataset. This is called *unsupervised* person re-ID. Roughly speaking, the current trend is to use a pre-trained base architecture to extract visual features, cluster them, and use the cluster assignments as *pseudo-labels* to re-train the base architecture using standard supervised re-ID loss functions [8], [9].

In parallel, since generative adversarial networks were proposed, adversarial learning has gained popularity in the domain adaptation field [10]–[12]. The underlying intuition is that learning a feature generator robust to the domain shift between *source* and *target* would improve the target performance. The adversarial learning paradigm has been successfully used for person re-ID in both the supervised [13], [14], and the unsupervised [7], [15] learning paradigms.

In this paper, we propose to unify these two trends in unsupervised person re-ID: hence using conditional adversarial networks for unsupervised person re-ID. Our intuition is that good person re-ID visual features should be independent of the

camera/viewpoint, see Fig. 1. Naturally, one would expect that an adversarial game between a generator (feature extractor) and a discriminator (camera classifier) should suffice. However, because the ID presence is not uniform in all cameras, such simple strategy implies some negative transfer and limits – often decreases – the representational power of the visual feature extractor. To overcome this issue, we propose to use conditional adversarial networks, thus providing an additional identity representation to the camera discriminator. Since in the target dataset, the ID labels are unavailable, we exploit the pseudo-labels. More precisely, we provide, as conditioning vector, the centroid of the cluster to which the image belongs. The contributions of this paper are the following:

- We investigate the impact of a camera-adversarial strategy in the unsupervised person re-ID task.
- We realize the negative transfer effect, and propose to use conditional adversarial networks.
- The proposed method can be easily plugged into any unsupervised clustering-based person re-ID methods. We experimentally combine **CANU** with two clustering-based unsupervised person re-ID methods, and propose to use their cluster centroids as conditioning labels.
- Finally, we perform an extensive experimental validation on four different unsupervised re-ID experimental settings and outperform current state-of-the-art methods by a large margin on all settings.

The rest of the paper is organized as follows. Section II describes the state-of-the-art. Section III discusses the basics of clustering-based unsupervised person re-ID and sets the notations. The proposed conditional adversarial strategy is presented in Section IV. The extensive experimental validation is discussed in Section V before drawing the conclusions in Section VI.

II. Related work

Unsupervised person re-identification (re-ID) has drawn growing attention in the last few years, taking advantage of the recent achievements of supervised person re-ID models, without requiring an expansive and tedious labeling process of the target data set. A very important line of research starts from a pre-trained model on the source data set and is based on *clustering* and *fine-tuning* [7]–[9], [15], [16]. It alternates between a clustering step generating noisy pseudo-labels, and a fine-tuning step adapting the network to the target data set distribution, leading to a progressive label refinement. Thus, these methods do not use the source data set during adaptation. A lot of effort has been invested in improving the quality of the pseudo-labels. Sampling from reliable clusters during adaptation [7], gradually reducing the number of clusters and merging by exploiting intrinsic inter-ID diversity and intra-ID similarity [15], or performing multiple clustering on visual sub-domains and enforcing consistency [8] have been investigated. More recently, [9] investigated the interaction of two different models to assess and incorporate pseudo-label reliability within a teacher-student framework.

A different approach is directly inspired by Unsupervised Domain Adaptation (UDA) [6], [17]–[21]: using both the source and target data sets during adaptation. These methods aim to match the distributions on the two data sets while keeping its discriminative ability leveraging source ground truth ID labels. A first strategy learns to map source's detections to target's style detections, and train a re-ID model in a supervised setting using those only those transferred detections [6], or in combination with the original target detections [17]. More standard UDA strategies use adversarial learning to match the source and target distributions [11], [19].

Negative transfer has been investigated in unsupervised domain adaptation [22], especially for Partial Domain Adaptation (PDA) [23]–[25], where target labels are only a subset of the source's. Negative transfer is defined as the inability of an adaptation method to find underlying common representation between data sets and is generally caused by the gap between the distributions of the two data sets being too wide [26] for the algorithm to transfer knowledge. Weighting mechanisms are generally employed to remove the impact of source's outliers class on the adaptation process, either for the matching part [24], [25], [27], the classification part [26], or both [23]. Interestingly, [26] uses a domain discriminator conditioned by source label to perform conditional distribution matching. Investigating negative transfer is not limited to UDA settings. For example, a similar method has been proposed for domain generalization [28], implementing a conditional discriminator to match conditioned domain distributions. By doing so, the impact of the difference between prior label distributions on the discriminative ability of the model is alleviated.

Within the task of unsupervised person re-ID, different cameras could be considered as different domains, and standard matching strategies could be used. However, they would inevitably induce negative transfer as described before for generic domain adaptation. Direct application of PDA methods into the person re-ID tasks is neither simple nor expected to be successful. The main reason is that, while PDA methods handle a few dozens of classes, standard re-ID data sets contain a few thousands of IDs. This change of scale requires a different strategy, and we propose to use conditional adversarial networks, with a conditioning label that describes the average sample in the cluster, rather than representing the cluster index. In conclusion, different from clustering and fine-tuning unsupervised person re-ID methods, we propose to exploit (conditional) adversarial networks to learn visual features that are camera independent and thus more robust to appear changes. Different from previous domain adaptation methods, we propose to match domains (cameras) with a conditioning label that evolves during training, since it is the centroid of the cluster to which the visual sample is assigned, allowing us having a representation that is independent of the number of clusters and the cluster index.

III. Clustering based Unsupervised Person Re-ID

We propose to combine conditional adversarial networks with clustering-based unsupervised person Re-ID. To detail

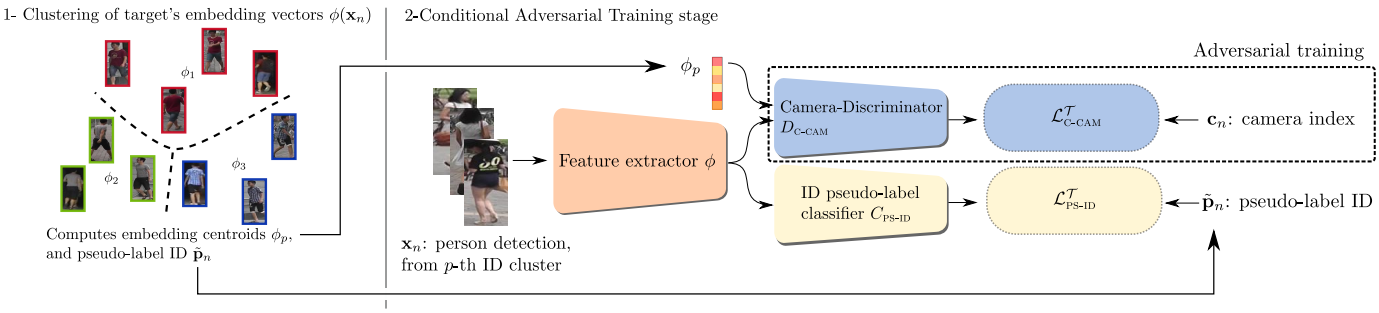


Fig. 2. Pipeline of our method: alternatively (1) clustering target's training data set using ϕ representation, producing noisy pseudo-label ID $\tilde{p}_n$ alongside centroids ϕ_p, and (2) conditional adversarial training, using a Camera-Discriminator D_{CAM} conditioned by ϕ_p to enforce camera invariance on a per identity basis to avoid negative transfer. Pseudo-label ID are used to train an ID classifier C_{PS-ID} alongside the discriminator.

our contributions, we first set up the basics and notations of existing methods for unsupervised person re-ID.

Let $\mathcal{S}$ denote a source ID-annotated person re-ID dataset, containing $N^{\mathcal{S}}$ images corresponding to $M^{\mathcal{S}}$ different person identities captured by $K^{\mathcal{S}}$ cameras. We write $\mathcal{S} = \{(\mathbf{x}_n^{\mathcal{S}}, \mathbf{p}_n^{\mathcal{S}}, \mathbf{c}_n^{\mathcal{S}})\}_{n=1}^{N^{\mathcal{S}}}$, where each three-tuple consists of a detection image, $\mathbf{x}_n^{\mathcal{S}}$, a person ID one-hot vector, $\mathbf{p}_n^{\mathcal{S}} \in \{0,1\}^{M^{\mathcal{S}}}$ and a camera index one-hot vector, $\mathbf{c}_n^{\mathcal{S}} \in \{0,1\}^{K^{\mathcal{S}}}$. Similarly, we define $\mathcal{T} = \{(\mathbf{x}_n^{\mathcal{T}}, \mathbf{c}_n^{\mathcal{T}})\}_{n=1}^{N^{\mathcal{T}}}$ a target person re-ID dataset, with $K^{\mathcal{T}}$ cameras and $N^{\mathcal{T}}$ element, without ID labels.

Source pre-training Let ϕ be a convolutional neural network backbone (*e.g.* ResNet-50 [29]) served as a trainable *feature extractor*. The goal of person re-ID is to be able to discriminate person identities, and therefore an identity classifier C_{ID} is required. The output of C_{ID} is a $M^{\mathcal{S}}$-dimensional stochastic vector, encoding the probability of the input to belong to each of the identities. The cross-entropy and triplet losses are usually employed:

$$\mathcal{L}^{\mathcal{S}}_{\text{CE}}(\phi, C_{\text{ID}}) = -\mathbb{E}_{(\mathbf{x}^{\mathcal{S}}, \mathbf{p}^{\mathcal{S}})\sim\mathcal{S}} \left\{ \log \left\langle C_{\text{ID}}(\phi(\mathbf{x}^{\mathcal{S}})), \mathbf{p}^{\mathcal{S}} \right\rangle \right\}, \quad (1)$$

$$\mathcal{L}^{\mathcal{S}}_{\text{TRI}}(\phi) = \mathbb{E}_{(\mathbf{x}^{\mathcal{S}}, \mathbf{x}_p^{\mathcal{S}}, \mathbf{x}_n^{\mathcal{S}})\sim\mathcal{P}_{\mathcal{S}}} \{\max(0, \|\phi(\mathbf{x}^{\mathcal{S}}) - \phi(\mathbf{x}_p^{\mathcal{S}})\| + m - \|\phi(\mathbf{x}^{\mathcal{S}}) - \phi(\mathbf{x}_n^{\mathcal{S}})\|)\}, \quad (2)$$

where $\mathbb{E}$ denotes the expectation, $\langle \cdot, \cdot \rangle$ the scalar product, $\|.\|$ the L^2-norm distance, $\mathbf{x}_p^{\mathcal{S}}$ and $\mathbf{x}_n^{\mathcal{S}}$ are the hardest positive and negative example for $\mathbf{x}^{\mathcal{S}}$ in $\mathcal{P}_{\mathcal{S}}$ the set of all triplets in $\mathcal{S}$, and $m = 0.5$. We similarly denote $\mathcal{L}^{\mathcal{T}}_{\text{CE}}$ and $\mathcal{L}^{\mathcal{S}}_{\text{TRI}}$ the cross-entropy and triplet losses evaluated on the target dataset. However, in unsupervised reID settings, target ID labels are unavailable, and therefore we will need to use alternative *pseudo-ID labels*. The re-ID feature extractor ϕ is typically trained using:

$$\mathcal{L}^{\mathcal{S}}_{\text{ID}}(\phi, C_{\text{ID}}) = \mathcal{L}^{\mathcal{S}}_{\text{CE}}(\phi, C_{\text{ID}}) + \lambda \mathcal{L}^{\mathcal{S}}_{\text{TRI}}(\phi), \quad (3)$$

for a fixed balancing value λ, achieving competitive performance on the source test set [30]. However, they notoriously lack generalization power and perform badly on datasets unseen during training [6], thus requiring adaptation.

Target fine-tuning As discussed above, target ID labels are unavailable. To overcome this while leveraging the discriminative power of widely-used losses described in Eq. 3, methods like [8], [9] use pseudo-labels. The hypothesis of these methods is that the features learned during the pre-training stage are exploitable for the inference of target's ID labels to a certain extent. Starting from the pre-trained model, these methods alternate between (i) pseudo ID label generation $\{\tilde{\mathbf{p}}_n^{\mathcal{T}}\}_{n=1}^{N^{\mathcal{T}}}$ using a standard clustering algorithm (k-means or DBSCAN [31]) on the target training set $\{\phi(\mathbf{x}_n^{\mathcal{T}})\}_{n=1}^{N^{\mathcal{T}}}$ and (ii) the update of ϕ using losses similar to Eq. 3 supervised by $\{\tilde{\mathbf{p}}_n^{\mathcal{T}}\}_{n=1}^{N^{\mathcal{T}}}$. Since our approach is agnostic to the ID loss used at this step, we choose to denote it by $\mathcal{L}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}})$, $C_{\text{PS-ID}}$ being an optional classifier layer for the pseudo-labels, and develop it further in the experimental section.

IV. CANU-ReID: A Conditional Adversarial Network for Unsupervised Person re-ID

In this section we discuss the main limitation of clustering-based unsupervised re-ID methods: we hypothesize that viewpoint variability can make things difficult for clustering methods and propose two alternatives. First, an adversarial network architecture targeting re-ID features that are camera-independent. This strategy could, however, induce some negative transfer when the correlation between cameras and IDs is strong. Second, a conditional adversarial network architecture specifically designed to overcome this negative transfer.

Camera adversarial-guided clustering We hypothesize that camera (viewpoint) variability is one of the major limiting factors for clustering-based unsupervised re-ID methods. In plain, if the embedding space variance explained by camera changes is high, the clustering method could be clustering images from the same camera, rather than images from the same ID. Therefore, ϕ will produce features that can very well discriminate the camera at the expense of the ID. To alleviate this problem, we propose to directly enforce camera invariance in ϕ's representation by using an adversarial strategy, where the discriminator is trained to recognize the camera used to capture the image. Consequently, the generator, in our case ϕ, is trained to remove any trace from the camera index (denoted by $\mathbf{c}$). Intuitively, this should reduce the viewpoint variance in the embedding space, improve pseudo-labels quality and increase the generalization ability of ϕ to unseen IDs.

To do so, we require a camera discriminator D_{CAM} (see Fig. 2 for a complete overview of the architecture). The generator ϕ and the discriminator D_{CAM} will be trained through a min-max formulation:

$$\min_{\phi, C_{\text{PS-ID}}} \max_{D_{\text{CAM}}} \mathcal{L}^{\mathcal{T}}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}}) - \mu \mathcal{L}^{\mathcal{T}}_{\text{CAM}}(\phi, D_{\text{CAM}}), \quad (4)$$

where $\mu > 0$ is a balance hyper-parameter that can be interpreted as a regularization parameter [11], and $\mathcal{L}^{\mathcal{T}}_{\text{CAM}}$ is defined via the cross-entropy loss:

$$\mathcal{L}^{\mathcal{T}}_{\text{CAM}}(\phi, D_{\text{CAM}}) = -\mathbb{E}_{(\mathbf{x}^{\mathcal{T}}, \mathbf{c}^{\mathcal{T}}) \sim \mathcal{T}} \left\{ \log \left\langle D_{\text{CAM}}(\phi(\mathbf{x}^{\mathcal{T}})), \mathbf{c}^{\mathcal{T}} \right\rangle \right\} \quad (5)$$

On one side, the feature extractor ϕ must minimize the person re-ID loss $\mathcal{L}_{\text{PS-ID}}$ at the same time as making the problem more challenging for the camera discriminator. On the other side, the camera discriminator tries to learn to recognize the camera corresponding to the input image.

Adversarial negative transfer It has been shown [28] that minimizing (4) is equivalent to the following problem:

$$\min_{\phi, C_{\text{PS-ID}}} \mathcal{L}^{\mathcal{T}}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}}) \quad (6)$$
$$\text{s.t. } \text{JSD}_{\mathcal{T}}(p(\phi(\mathbf{x})|\mathbf{c}=1), \dots, p(\phi(\mathbf{x})|\mathbf{c}=K)) = 0,$$

where $\text{JSD}_{\mathcal{T}}$ stands for the multi-distribution Jensen-Shanon divergence [32] on the target set $\mathcal{T}$, and we dropped the superscript $\mathcal{T}$ in the variables to ease the reading.

Since the distribution of ID labels may strongly depend on the camera, the plain adversarial strategy in (6) can introduce negative transfer [26]. Formally, since we have:

$$p(\mathbf{p}|\mathbf{c}=i) \neq p(\mathbf{p}|\mathbf{c}=j), i \neq j$$

then solving (6) is not equivalent (see [28]) to:

$$\min_{\phi, C_{\text{PS-ID}}} \mathcal{L}^{\mathcal{T}}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}}) \quad (7)$$
$$\text{s.t. } \text{JSD}_{\mathcal{T}}(p(\phi(\mathbf{x})|\mathbf{p}, \mathbf{c}=1), \dots, p(\phi(\mathbf{x})|\mathbf{p}, \mathbf{c}=K)) = 0,$$

which is the problem we would implicitly want to solve. Intuitively, *negative transfer* means that the camera discriminator learns $p(\mathbf{c}|\mathbf{p})$ instead of $p(\mathbf{c}|\mathbf{x}, \mathbf{p})$, exploiting ID to infer camera information and decreasing the representation power of ϕ due to the adversarial loss.

Conditional adversarial networks We propose to directly solve the optimization problem in Eq. 7 to alleviate the negative transfer. Similar to the original conditional GAN formulation [33], we condition the adversarial discriminator with the input ID $\mathbf{p}$. Given that ID labels are unavailable on the target set, we replace them by the pseudo-labels obtained during the clustering phase.

However, since we are handling a large number of IDs (700 to 1500 in standard re-ID datasets), using a one-hot representation turned out to be very ineffective. Indeed, such representation is not permutation-invariant, meaning that if the clusters are re-ordered, the associated conditional vector changes, which does not make sense. We, therefore, need a permutation-invariant conditioning label.

To do so, we propose to use the cluster centroids $\phi_{\mathbf{p}}$ which are provided by the clustering algorithms at no extra cost. This conditioning vectors are permutation invariant. Importantly, we do not back-propagate the adversarial loss through the ID-branch, to avoid using an ID-dependant gradient from the adversarial loss. This boils down to defining $\mathcal{L}_{\text{C-CAM}}$ as:

$$\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi, D_{\text{C-CAM}}) = -\mathbb{E}_{(\mathbf{x}, \mathbf{p}, \mathbf{c}) \sim \mathcal{T}} \left\{ \log \left\langle D_{\text{C-CAM}}(\phi(\mathbf{x}), \phi_{\mathbf{p}}), \mathbf{c} \right\rangle \right\} \quad (8)$$

and then solving:

$$\min_{\phi, C_{\text{PS-ID}}} \max_{D_{\text{C-CAM}}} \mathcal{L}^{\mathcal{T}}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}}) - \mu \mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi, D_{\text{C-CAM}}). \quad (9)$$

V. Experimental Validation

In this section, we provide implementation details and an in-depth evaluation of the proposed methodology, setting the new state-of-the-art in four different unsupervised person re-ID experimental settings. We also provide an ablation study and insights on why conditional adversarial networks outperform existing approaches.

A. Evaluation Protocol

We first describe here the baselines, on which our proposed **CANU** is built and tested. The used datasets and the evaluation metrics are then introduced.

Baselines The proposed **CANU** can be easily plugged into any clustering-based unsupervised person re-ID methods. Here, we experimentally test it on two state-of-the-art clustering-based unsupervised person re-ID methods, as baselines.

First, self-similarity grouping [8] (**SSG**) performs independent clustering on the upper-, lower- and full-body features, denoted as ϕ^{U}, ϕ^{L} and ϕ^{F}. They are extracted from three global average pooling layers of the convolutional feature map of ResNet-50 [29]. The underlying hypothesis is that noisy global pseudo-label generation can be improved by using multiple, but related clustering results, and enforcing consistency between them. The triplet loss is used to train the overall architecture.

To implement **CANU**-SSG, we define three different camera discriminators, one for each embedding, $D^{\text{U}}_{\text{C-CAM}}$, $D^{\text{L}}_{\text{C-CAM}}$ and $D^{\text{F}}_{\text{C-CAM}}$ respectively, each fed with samples from the related representation and conditioned by the global embedding ϕ^{F}. In the particular case of **CANU**-SSG, the generic optimisation problem in Eq. 9 instantiates as:

$$\min_{\phi} \max_{D^{\text{U,L,F}}_{\text{C-CAM}}} \mathcal{L}^{\mathcal{T}}_{\text{SSG}}(\phi) - \mu \mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^{\text{U}}, D^{\text{U}}_{\text{C-CAM}}) \quad (10)$$
$$- \mu \mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^{\text{L}}, D^{\text{L}}_{\text{C-CAM}}) - \mu \mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^{\text{F}}, D^{\text{F}}_{\text{C-CAM}}).$$

Second, Mutual Mean-Teaching [9] (**MMT**) reduces pseudo-label noise by using a combination of hard and soft assignment: using hard labeling reduces the amount of information given to the model, and using soft labeling allows the cluster's confidence to be taken into account. MMT defines two different models $(\phi^1, C^1_{\text{PS-ID}})$ and $(\phi^2, C^2_{\text{PS-ID}})$, both implemented with a IBN-ResNet-50 [34] backbone, initialized with

two different pre-trainings on the source dataset. They are then jointly trained using pseudo labels as hard assignments, and inspired by teacher-student methods, using their own pseudo ID predictions as soft pseudo-labels to supervise each other. Soft versions of cross-entropy and triplet loss are used.

To implement **CANU**-MMT, similar to **CANU**-SSG, we define two camera discriminators $D^1_{\text{C-CAM}}$ and $D^2_{\text{C-CAM}}$, each dedicated to one embedding, and train it using the following instantiation of the generic optimisation problem in Eq. 9:

$$\min_{\phi^{1,2}, C^{1,2}_{\text{PS-ID}}} \max_{D^{1,2}_{\text{C-CAM}}} \mathcal{L}^{\mathcal{T}}_{\text{MMT}}(\phi^1, C^1_{\text{PS-ID}}) + \mathcal{L}^{\mathcal{T}}_{\text{MMT}}(\phi^2, C^2_{\text{PS-ID}}) \\ - \mu\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^1, D^1_{\text{C-CAM}}) - \mu\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^2, D^2_{\text{C-CAM}}). \tag{11}$$

While the clustering strategy used in SSG is DBSCAN [31], the one used in MMT is standard k-means. For a fair comparison, we implemented **CANU** with DBSCAN, which has the advantage of automatically selecting the number of clusters. We also evaluate the performance of MMT using the DBSCAN clustering strategy without **CANU**, to evaluate the impact of our method on a fair basis.

Datasets The proposed adversarial strategies are evaluated using three datasets: Market-1501 (Mkt) [5], DukeMTMC-reID (Duke) [4] and MSMT17 (MSMT) [35]. In all three cases, the dataset is divided into three parts: training, gallery, and query. The query and the gallery are never available during training and only used for testing.

Mkt is composed of $M = 1,501$ (half for training and half for testing) different identities, observed through $K = 6$ different cameras (viewpoints). The deformable parts model [36] is used for person detection. As a consequence, there are $N = 12,936$ training images and $19,732$ gallery images. In addition, there are $3,368$ hand-drawn bounding box queries.

Duke is composed of $M = 1,404$ (half for training and half for testing) identities captured from $K = 8$ cameras. In addition, 408 other ID, called "distractors", are added to the gallery. Detections are manually selected, leading to $N = 16,522$ images for train, $17,661$ for the gallery and $2,228$ queries.

MSMT is the largest and most competitive dataset available, with $M = 4,101$ identities ($1,041$ for training, and $3,060$ for test), $K = 15$ cameras, with $N = 32,621$ images for training, $82,161$ for the Gallery and $11,659$ queries.

The unsupervised person re-ID experimental setting using dataset A as source and dataset B as the target is denoted by A ► B. We compare the proposed methodology in four different settings: Mkt ► Duke, Duke ► Mkt, Mkt ► MSMT and Duke ► MSMT.

Evaluation metrics In order to provide an objective evaluation of the performance, we employ two standard metrics in person re-ID [5]: Rank-1 (R1) and mean average-precision (mAP). Precisely, for each query image, we extract visual features employing ϕ, and we compare them to the features extracted from the gallery using the cosine distance. Importantly, the gallery images captured with the same camera as the query image are not considered. For R1, a query is well identified if the closest gallery feature vector corresponds to the same identity. In the case of mAP, the whole list of gallery images is considered, and precision at different ranking positions is averaged. See [5] for details. For both metrics, the mean over the query set is reported.

Implementation details For both MMT and SSG, we use the models pre-trained on the source datasets (e.g. For Mkt ►Duke, we use the model pre-trained on the Market dataset and provided by [8] and [9]). DBSCAN is used at the beginning of each training epoch, the parameters for DBSCAN are the same described as in [8]. The weight for (conditional) adversarial losses μ is set to 0.1 for MMT and to 0.05 for SSG, chosen according to a grid search with values between $[0.01, 1.8]$ (see below). The used conditional discriminator has two input branches, one as the (conditional) ID branch and the other is the camera branch, both consist of four fully-connected layers, of size $[2048, 1024]$, $[2048, 1024]$, $[1024, 1024]$, $[1024, \text{number of cameras}]$, respectively. Batch normalization [37] and ReLU activation are used. For MMT, during the unsupervised learning, we train the IBN-ResNet-50 [34] feature extractor with Adam [38] optimizer using a learning rate of 0.00035. As default in [9], the network is trained for 40 epochs but with fewer iterations per epoch (400 v.s. 800 iterations) while keeping a similar or better performance. For SSG, we train the ResNet-50 [29] with SGD optimizer using a learning rate of $6e$-5. At each epoch, unlike MMT, we iterate through the whole training set instead of training with a fix number of iterations.

After training, the discriminator is discarded and only the feature extractor is kept for evaluations. For SSG, first, it combines the features extracted from the original image and the horizontally flipped image with a simple sum. Second, the summed features are normalized by their L_2 norm. Finally, The full-, upper- and, lower-body normalized features are concatenated to form the final features. For MMT, the features extracted from the feature extractor are directly used for evaluations.

In the following, we first compare the proposed methodology with the state-of-the-art (see Sec. V-B). Secondly, we discuss the benefit of using conditional camera-adversarial training in the ablation study (see Sec. V-C), and include several insights on the performance of **CANU**.

B. Comparison with the State-of-the-Art

We compare **CANU**-SSG and **CANU**-MMT to the state-of-the-art methods and we demonstrate in Tables I and II that **CANU**-MMT sets a new state-of-the-art result compared to the existing unsupervised person re-ID methods by a large margin. In addition, **CANU**-SSG outperforms SSG in all settings. Since the MSMT dataset is more recent, fewer comparisons are available in the experiments involving this dataset, hence the two different tables.

More precisely, the proposed **CANU** significantly improves the performance of the baselines, SSG [8] and MMT [9]. In Mkt ►Duke and Duke ►Mkt (Table I), **CANU**-SSG improves SSG by ↑3.1%/↑3.6% (R1/mAP, same in the following.)

TABLE I
COMPARISON OF THE PROPOSED **CANU** METHODOLOGY ON THE MKT ▶ DUKE AND DUKE ▶ MKT UNSUPERVISED PERSON RE-ID SETTINGS. **CANU-MMT** ESTABLISHES A NEW STATE-OF-THE-ART IN BOTH SETTINGS, AND **CANU-SGG** OUTPERFORMS **SSG**.

Method	Mkt ▶ Duke		Duke ▶ Mkt	
	R1	mAP	R1	mAP
PUL [7]	30.0	16.4	45.5	20.5
TJ-AIDL [39]	44.3	23.0	58.2	26.5
SPGAN [6]	41.1	22.3	51.5	22.8
HHL [17]	46.9	27.2	62.2	31.4
CFSM [18]	49.8	27.3	61.2	28.3
BUC [15]	47.4	27.5	66.2	38.3
ARN [40]	60.2	33.4	70.3	39.4
UDAP [20]	68.4	49.0	75.8	53.7
ENC [21]	63.3	40.4	75.1	43.0
UCDA-CCE [19]	47.7	31.0	60.4	30.9
PDA-Net [41]	63.2	45.1	75.2	47.6
PCB-PAST [16]	72.4	54.3	78.4	54.6
Co-teaching [42]	77.6	61.7	87.8	71.7
SSG [8]	73.0	53.4	80.0	58.3
CANU-SSG (ours)	76.1	57.0	83.3	61.9
MMT [9]	81.8	68.7	91.1	74.5
MMT (DBSCAN)	80.2	67.2	91.7	79.3
CANU-MMT (ours)	**83.3**	**70.3**	**94.2**	**83.0**

TABLE II
COMPARISON OF THE PROPOSED **CANU** METHODOLOGY ON THE MKT ▶ MSMT AND DUKE ▶ MSMT UNSUPERVISED PERSON RE-ID SETTINGS. **CANU-MMT** ESTABLISHES A NEW STATE-OF-THE-ART IN BOTH SETTINGS, AND **CANU-SGG** OUTPERFORMS **SSG**.

Method	Mkt ▶ MSMT		Duke ▶ MSMT	
	R1	mAP	R1	mAP
PTGAN [43]	10.2	2.9	11.8	3.3
ENC [21]	25.3	8.5	30.2	10.2
SSG [8]	31.6	13.2	32.2	13.3
CANU-SSG (ours)	45.5	19.1	43.3	17.9
MMT [9]	54.4	26.6	58.2	29.3
MMT (DBSCAN)	51.6	26.6	59.0	32.0
CANU-MMT (ours)	**61.7**	**34.6**	**66.9**	**38.3**

TABLE III
IMPACT OF μ IN THE PERFORMANCE OF **CANU**. WHEN THE MAP VALUES ARE EQUAL, WE HIGHLIGHT THE ONE CORRESPONDING TO HIGHER R1.

Method	μ	Mkt ▶ Duke		Duke ▶ Mkt	
		R1	mAP	R1	mAP
CANU-SSG	0.01	72.8	53.3	79.7	57.2
	0.05	**76.1**	**57.0**	**83.3**	**61.9**
	0.1	74.7	56.2	82.7	61.1
	0.2	75.3	56.5	81.8	60.3
	0.4	73.3	53.5	80.4	59.2
	1.8	7.1	2.9	39.1	17.1
CANU-MMT	0.01	81.3	68.9	92.6	79.2
	0.05	82.4	70.3	93.0	81.3
	0.1	**83.3**	**70.3**	**94.2**	**83.0**
	0.2	82.7	70.3	93.4	82.5
	0.4	82.5	70.3	93.8	82.0
	1.8	82.8	69.9	93.1	81.3

and ↑3.3%/↑3.6% respectively, and **CANU**-MMT significantly outperforms MMT by ↑1.5%/↑1.6% and ↑3.1%/↑8.5% respectively. Moreover, for the more challenging setting (Table II), the improvement brought by **CANU** is even more evident. For SSG, for example, we increase the R1/mAP by ↑13.9%/↑5.9% in Mkt ▶ MSMT, and by ↑11.1%/↑4.6% in Duke ▶ MSMT. For MMT, **CANU**-MMT outperforms MMT by ↑7.3%/↑8.0% in Mkt ▶ MSMT, and by ↑8.7%/↑9.0% in Duke ▶ MSMT. Finally, the consistent improvement in the four settings of **CANU**-MMT over MMT (DBSCAN) and the inconsistent improvement of MMT (DBSCAN) over standard MMT proves that the increase of the performance is due to the proposed methodology. To summarize, we greatly improve the baselines using the proposed **CANU**. More importantly, to our best knowledge, we outperform the existing methods by a large margin and establish a new state-of-the-art result.

C. Ablation Study

In this section, we first perform a study to evaluate the impact of the value of μ. Secondly, we demonstrate the interest of the conditional strategy, versus its non-conditional counterpart. Thirdly, we study the evolution of the mutual information between ground-truth camera indexes and pseudo-labels using MTT (DBSCAN), thus providing some insights on the quality of the pseudo-labels and the impact of the conditional strategy on it. Finally, we visualize the evolution of the number of lost person identities at each training epoch, to assess the impact of the variability of the training set.

Selection of μ We ablate the value μ by comparing the performance (R1 and mAP) of models trained within the range $[0.01, 1.8]$. From Tab. III, $\mu = 0.1$ (**CANU**-MMT) and $\mu = 0.05$ (**CANU**-SSG) yield the best person re-ID performance.

Is conditional necessary? From Table IV, we show that the camera adversarial network can help the person re-ID networks trained with clustering-based unsupervised methods better capture the person identity features: **CANU** and adding a simple adversarial discriminator (+Adv.) significantly outperform the baseline methods in all settings. This is due to the combination of the camera adversarial network with unsupervised clustering-based methods. By doing so, the camera dependency is removed from the features of each person thus increasing the quality of the overall clustering. However, because of the negative transfer effect, the camera adversarial network cannot fully exploit the camera information while discarding the person ID information. For this reason, the

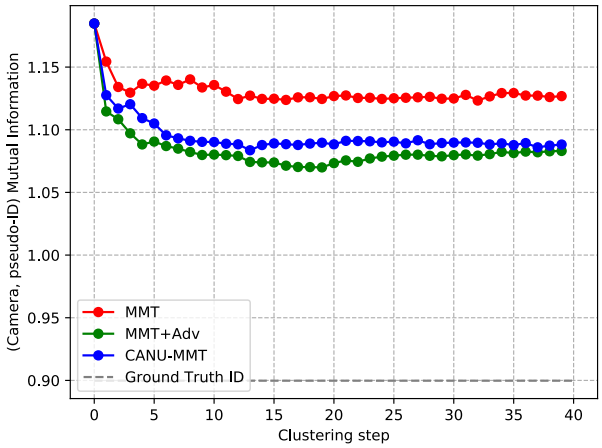
TABLE IV
EVALUATION OF THE IMPACT OF THE CONDITIONAL STRATEGY ON SGG [8] AND MMT [9] (USING DSCAN). WHEN THE MAP VALUES ARE EQUAL, WE HIGHLIGHT THE ONE CORRESPONDING TO HIGHER R1.

Method	Mkt ► Duke		Duke ► Mkt	
	R1	mAP	R1	mAP
SSG [8]	73.0	53.4	80.0	58.3
SSG+Adv.	75.4	56.4	**83.8**	**62.7**
CANU-SSG	**76.1**	**57.0**	83.3	61.9
MMT (DBSCAN)	80.2	67.2	91.7	79.3
MMT+Adv.	82.6	70.3	93.6	82.2
CANU-MMT	**83.3**	**70.3**	**94.2**	**83.0**

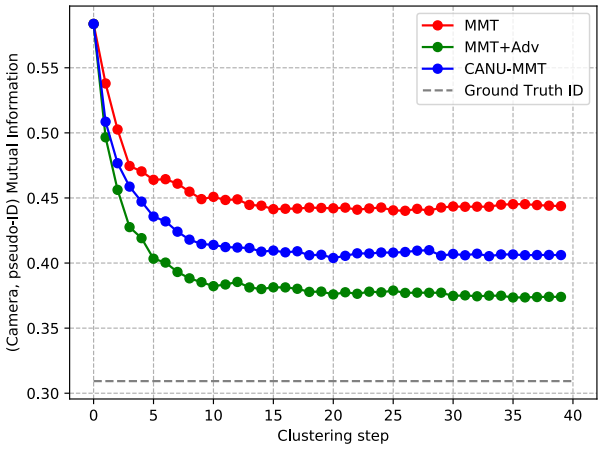
proposed method **CANU** improves the capacity of the camera adversarial network over the simple adversarial strategy. In summary, we demonstrate that the camera adversarial network can help improve the results of unsupervised clustering-based person re-ID. Moreover, the proposed **CANU** further improves the results by removing the link between camera and IDs.

Removing camera information Table IV demonstrates that removing camera information is globally positive, but that can also be harmful if it is not done with care. In this section, we further demonstrate that the proposed adversarial strategies actually reduce the camera dependency in clustering results and present some insights on why the conditional strategy is better than the plain adversarial network. To do so, we plot the mutual information between the pseudo-labels provided by DBSCAN, and the fixed camera index information, at each clustering stage (i.e. training epoch) in Fig. 3. Intuitively, the mutual information between two variables is a measure of mutual dependence between them: the higher it is, the more predictable one is from knowing the other. We report the results for MMT on Duke ► Mkt and Mkt ► Duke, **CANU**-MMT and the simple adversarial strategy. We observe that the mutual information is systematically decreasing with the training, even for plain MMT. Both adversarial strategies significantly outperform plain MMT at reducing the camera-pseudo-ID dependency, **CANU**-MMT being slightly less effective than MMT+Adv. This is consistent with our theoretical framework, since matching ID-conditioned camera distribution in ϕ does not account for the ID-Camera dependency, and thus is less effective in terms of camera dependency, but preserves identity information, see Table IV. We also observe that there is a significant gap between the target mutual information (i.e. measured between ground truth ID and camera index) for all methods, which exhibits the performance gap between supervised and unsupervised person re-ID methods.

Evolution of the number of lost IDs Since we train the target dataset using unsupervised techniques, we do not use the ground-truth labels in the target dataset during training. Instead, we make use of the pseudo labels provided by DBSCAN. DBSCAN discards the outliers i.e. features that are not closed to others. It is natural to wonder how many



(a) Mkt ► Duke



(b) Duke ► Mkt

Fig. 3. Mutual information between pseudo labels and camera index evolution for the MMT setting. Ground-truth ID comparison is displayed in dashed lines for both datasets.

identities are "lost" at every iteration. We here visualize the number of lost ID (all those that are not present in a training epoch) after each clustering step. We plot the evolution of this number with the training epoch for MMT, MMT+Adv. and **CANU**-MMT on Duke ► Mkt in Fig. 4. The dual experiment, i.e. on Mkt ► Duke revealed that no ID was lost by any method. In Fig. 4, we first observe that the loss of person identities decreases with the clustering steps. It means that the feature extractor provides more and more precise features representing person identities. Secondly, the use of camera adversarial training can reduce the loss of person identities in the clustering algorithm, which reflects the benefit of camera adversarial networks to the clustering algorithm and thus to the unsupervised person re-ID task.

VI. CONCLUSION

In this paper, we demonstrate the benefit of unifying adversarial learning with current unsupervised clustering-based person re-identification methods. We propose to condition the adversarial learning with the cluster centroids, being these

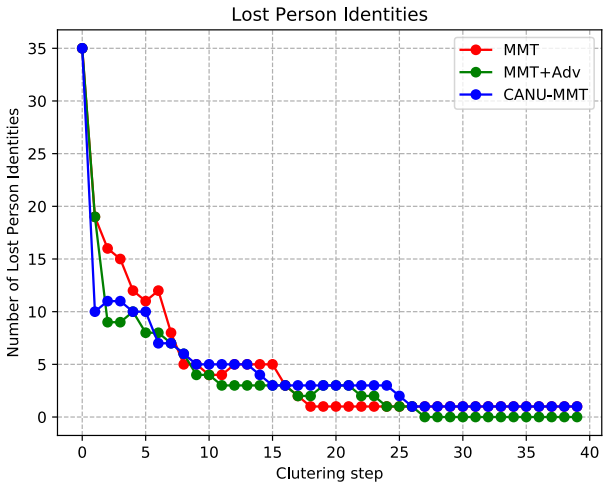


Fig. 4. Evolution of the number of lost person IDs during training using MMT on Duke ► Mkt.

representations independent of the number of clusters and invariant to cluster index permutations. The proposed strategy boosts existing clustering-based unsupervised person re-ID baselines and sets the new state-of-the-art performance in four different unsupervised person re-ID experimental settings. We believe that the proposed method **CANU** was a missing component in training unsupervised person re-identification networks and we hope that our work can give insight to this direction in the person re-identification domain.

REFERENCES

[1] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014.

[2] T. Matsukawa and E. Suzuki, "Person re-identification using cnn features learned from combination of attributes," in *ICPR*, 2017.

[3] N. Jiang, J. Liu, C. Sun, Y. Wang, Z. Zhou, and W. Wu, "Orientation-guided similarity learning for person re-identification," in *ICPR*, 2018.

[4] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV Workshops*, 2016.

[5] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE ICCV*, 2015.

[6] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *IEEE CVPR*, 2018.

[7] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2018.

[8] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *IEEE ICCV*, 2019.

[9] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," *ICLR*, 2020.

[10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *IEEE CVPR*, 2017.

[11] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, 2016.

[12] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *NIPS*, 2016.

[13] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *IEEE CVPR*, 2018.

[14] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *IEEE ICCV*, 2017.

[15] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI*, 2019.

[16] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *IEEE CVPR*, 2019.

[17] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *ECCV*, 2018.

[18] X. Chang, Y. Yang, T. Xiang, and T. M. Hospedales, "Disjoint label space transfer learning with common factorised space," in *AAAI*, 2019.

[19] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, and Y. Gao, "A novel unsupervised camera-aware domain adaptation framework for person re-identification," in *IEEE CVPR*, 2019.

[20] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognition*, 2020.

[21] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *IEEE CVPR*, 2019.

[22] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications*, E. Soria, J. Martin, R. Magdalena, M. Martinez, and A. Serrano, Eds., 2009.

[23] M. L. J. W. Q. Y. Zhangjie Cao, Kaichao You, "Learning to transfer examples for partial domain adaptation," in *IEEE CVPR*, June 2019.

[24] Z. Cao, M. Long, J. Wang, and M. I. Jordan, "Partial transfer learning with selective adversarial networks," in *IEEE CVPR*, 2018.

[25] J. Zhang, Z. Ding, W. Li, and P. Ogunbona, "Importance weighted adversarial nets for partial domain adaptation," *IEEE CVPR*, 2018.

[26] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," in *IEEE CVPR*, 2019, pp. 11 293–11 302.

[27] Y. Yao, Y. Zhang, X. Li, and Y. Ye, "Heterogeneous domain adaptation via soft transfer network," in *ACM MM*, 2019.

[28] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *ECCV*, 2018.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[30] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *ArXiv preprint*, 2017.

[31] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[32] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, 1991.

[33] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[34] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *ECCV*, 2018.

[35] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *IEEE CVPR*, 2018.

[36] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint*, 2015.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *IEEE CVPR*, 2018, pp. 2275–2284.

[40] Y.-J. Li, F.-E. Yang, Y.-C. Liu, Y.-Y. Yeh, X. Du, and Y.-C. Frank Wang, "Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification," in *IEEE CVPR Workshops*, 2018, pp. 172–178.

[41] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C. F. Wang, "Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation," in *IEEE ICCV*, 2019.

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

[43] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *IEEE CVPR*, 2018.