# Camera-Driven Representation Learning for Unsupervised Domain Adaptive Person Re-identification

Geon Lee[1] Sanghoon Lee[1] Dohyung Kim[1] Younghoon Shin[2] Yongsang Yoon [2] Bumsub Ham[1*]

[1]Yonsei University [2]Robotics Lab, Hyundai Motor Company

https://cvlab.yonsei.ac.kr/projects/CaCL

## Abstract

*We present a novel unsupervised domain adaption method for person re-identification (reID) that generalizes a model trained on a labeled source domain to an unlabeled target domain. We introduce a camera-driven curriculum learning (CaCL) framework that leverages camera labels of person images to transfer knowledge from source to target domains progressively. To this end, we divide target domain dataset into multiple subsets based on the camera labels, and initially train our model with a single subset (i.e., images captured by a single camera). We then gradually exploit more subsets for training, according to a curriculum sequence obtained with a camera-driven scheduling rule. The scheduler considers maximum mean discrepancies (MMD) between each subset and the source domain dataset, such that the subset closer to the source domain is exploited earlier within the curriculum. For each curriculum sequence, we generate pseudo labels of person images in a target domain to train a reID model in a supervised way. We have observed that the pseudo labels are highly biased toward cameras, suggesting that person images obtained from the same camera are likely to have the same pseudo labels, even for different IDs. To address the camera bias problem, we also introduce a camera-diversity (CD) loss encouraging person images of the same pseudo label, but captured across various cameras, to involve more for discriminative feature learning, providing person representations robust to inter-camera variations. Experimental results on standard benchmarks, including real-to-real and synthetic-to-real scenarios, demonstrate the effectiveness of our framework.*

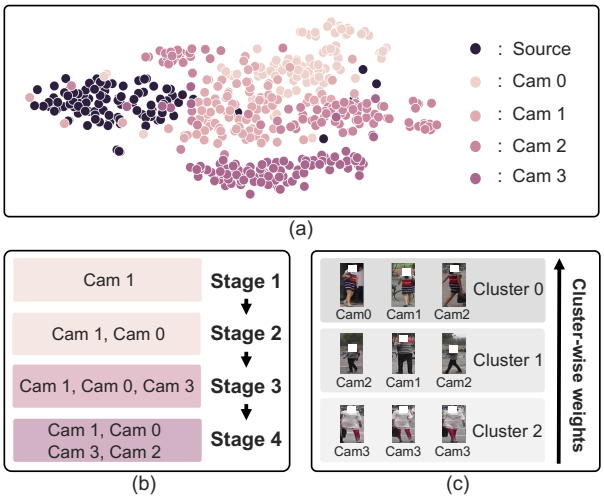


Figure 1: We visualize in (a) a t-SNE plot for features extracted from person images in Market1501 [55] and MSMT17 [45], using a reID model trained on MSMT17, where MSMT17 and Market1501 are source and target domains, respectively. The samples from different cameras in the target domain are distinguished by different colors. The model trained on a single domain offers features that are highly biased towards camera labels of person images for other domains. We propose to establish a camera-driven curriculum, as shown in (b), and initially train our model using images captured by a single camera, then gradually exploit more images captured using multiple cameras. To further alleviate the camera bias issue, we compute cluster-wise weights, as in (c), to encourage clusters containing images obtained from various cameras to involve more during the adaptation process.

## 1. Introduction

The objective of person re-identification (reID) is to retrieve person images of the same ID as a query person across non-overlapping cameras [48, 56]. Current reID approaches mainly adopt a supervised learning paradigm by exploiting person ID labels, and focus on learning discriminative person representations in a single domain. However, reID

*Corresponding author

models trained on a specific domain typically fail to generalize to other domains [7, 45], thus limiting the applicability in real-world scenarios. To address this issue, recent works [5, 45, 52, 53] exploit unsupervised domain adaptation techniques, transferring knowledge learned from a source domain to re-identify persons in a target one, where ID labels for the source domain are provided only [33]. This enables performing reID on the target domain without additional annotations, which is typically time-consuming and labor-intensive to obtain [29, 35, 55]. Unsupervised domain adaptive (UDA) reID is challenging due to the following reasons. Transferring knowledge from one domain to another is difficult due to the distribution gap between camera topologies for different domains [35, 45, 55]. Moreover, it is difficult to learn discriminative person representations for the target domain without ID labels, due to the large intra-class variations, particularly between person images captured by different cameras.

In recent years, most UDA reID methods [5, 11, 13, 19, 21, 31, 51, 52, 53, 57] exploit pseudo ID labels for target images to mitigate the discrepancies between source and target domains. To generate pseudo labels for the target domain, these methods first extract features from target images, using a reID model pre-trained on the source domain, and apply a clustering algorithm (*e.g.*, DBSCAN [9]) on the features. They then assign the same ID label to the images which belong to the same cluster, facilitating training with target images in a supervised manner. While the UDA reID methods have allowed significant advances for UDA reID, they mainly have two limitations. First, current approaches still focus on transferring knowledge from source to target in a domain-level. Namely, they attempt to adapt a model trained on a source domain to the target one *at once*, by regarding target images as a whole. This is not effective for transferring knowledge for UDA reID, since source and target domains have different camera topologies. Second, pseudo labels for the target domain are highly biased towards camera labels of images (Fig. 1(a)). That is, person images captured by the same camera are likely to be assigned to the same pseudo ID label, even for the persons with different IDs. Directly training a reID model with such labels rather hinders discriminative feature learning [5, 52, 53], particularly for the person images of the same ID but captured by different cameras.

In this paper, we present a novel framework for UDA reID that performs a *progressive* adaptation exploiting camera labels of person images. We conjecture that domain adaptation in a domain-level regime might be suboptimal, especially in the context of reID, since the distribution of a camera topology is highly unique for each domain. In order to consider an abrupt change on the camera topology from source to target domains, we propose a camera-driven curriculum learning (CaCL) leveraging camera labels of person images, facilitating a progressive adaptation (Fig. 1(b)). To implement this idea, we first decompose a target domain dataset into multiple subsets w.r.t the camera labels. Starting from a single subset (*i.e.*, images obtained from a single camera), we gradually add subsets to train our model, according to a curriculum sequence obtained by a camera-driven scheduling rule. The scheduler considers maximum mean discrepancies (MMD) [16] between each subset and the source domain dataset, such that a closer subset w.r.t the source dataset is exploited earlier within the curriculum. We also introduce a camera-diversity (CD) loss that encourages the clusters having person images obtained from various cameras to involve more for discriminative feature learning (Fig. 1(c)). It further incorporates a selective scheme for training that discards trivial clusters, only consisting of person images taken from the same camera. A model trained with CD loss is able to offer person representations more robust to inter-camera variations, compared to conventional cross-entropy [58] and triplet [20] losses, even when training with pseudo labels biased to camera labels. Together with CaCL and the CD loss, we achieve a new state of the art on standard UDA reID benchmarks, including real-to-real (*e.g.*, Market1501 [55]-to-MSMT17 [45] and MSMT17-to-Market1501) and synthetic-to-real (*e.g.*, PersonX [41]-to-Market1501 and Unreal [49]-to-MSMT17) scenarios, and demonstrate the effectiveness of our approach with extensive experimental results and ablative analyses.

Our main contributions can be summarized as follows: (1) We introduce a novel curriculum learning framework for UDA reID that leverages camera labels of person images. To the best of our knowledge, this is the first to incorporate a curriculum learning scheme for UDA reID. We also present the camera-driven scheduler that determines the curriculum sequence for multiple subsets in a target domain. (2) We present the CD loss to learn discriminative person representations, particularly robust to inter-camera variations, even when training with pseudo labels biased to camera labels. (3) We set a new state of the art on standard benchmarks for UDA reID, including real-to-real and synthetic-to-real scenarios, and demonstrate the effectiveness of our framework.

## 2. Related work

**UDA reID.** There are many attempts for UDA reID to handle the domain gap between source and target, without ID labels for the target domain, which can be categorized into two groups. The first line of works [7, 45] use generative models to translate person images of the source domain into the target one. Taking source images as input, target-stylized person images are generated using generative adversarial networks (GANs) [15] for image translation [61], with identity-preserving techniques [7]. The generated images are then used to train a reID model

on the target domain in a supervised manner. These approaches to exploiting generative models typically involve many heuristics [37], and require a lot of parameters, due to the unstable training of GANs [25]. Another line of works [5, 11, 13, 19, 21, 31, 51, 52, 53, 57] transfer knowledge from source to target domains using a self-training scheme [10]. These methods pre-train a reID model on the source domain with ID labels, and exploit the model to extract person representations from target images. They apply a clustering algorithm on the representations, and person images within the same cluster are assigned the same pseudo ID label. As the quality of pseudo labels largely influences the reID performance on the target domain, the works of [11, 51, 52, 57] attempt to refine the pseudo labels, *e.g.*, by leveraging multiple reID models and measuring prediction consistencies [51, 52]. We have observed that the person representations for target images, obtained using the source-pretrained model, offer clustering results that are highly biased towards camera labels. In this context, camera labels of person images can provide complementary information to alleviate this problem, which has not been considered previously. Moreover, all the aforementioned approaches do not consider the large discrepancies between camera topologies for different domains. Therefore, they handle the domain gap in a domain-level regime, considering the target domain as a whole. Instead of mitigating the domain gap at once, a recent approach [5] proposes to generate person representations of intermediate domains to perform adaptation gradually, by mixing cross-domain features [42]. This approach, however, still exploits all target images jointly during the adaptation process. In contrast to this, we start training with a subset of person images in the target domain, and gradually expand to multiple subsets, through a camera-driven curriculum for a progressive adaptation.

**ReID with auxiliary supervision.** Person reID methods focus on extracting discriminative person representations to match person of the same ID effectively, while differentiating persons of different IDs. Since it is challenging to handle large intra-class variations (*e.g.*, background clutter, viewpoint, and pose variations) with ID labels alone, many works exploit auxiliary supervisory signals for reID. Examples of the auxiliary signals include human pose [3, 12, 54], semantic parsing [22], and attribute labels [32, 38]. These provide additional cues for, *e.g.*, a part-to-part matching [22] or a feature disentanglement [3], enhancing the discriminative power of person representations. The auxiliary labels are expensive to obtain, as they use additional networks [3, 28, 46] trained with task-specific datasets [1, 14] or require labor-intensive annotations [8, 26]. Camera labels of input images, on the other hand, provide an efficient alternative, since they can easily be accessed from the metadata of images [60]. There are attempts to leverage camera labels of input images during training, to learn person representations robust to inter-camera variations, which is particularly important for reID that performs person matching in a cross-camera setting. For example, the work of [62] computes camera-specific feature statistics to mitigate the distribution gap between different cameras. In the context of UDA reID, the work of [34] proposes to use a discriminator for camera labels within an adversarial learning framework. The work of [60] generates multiple images of the same person in the style of different cameras, and uses the synthesized person images for UDA reID. In contrast to the UDA reID approaches to exploiting camera labels [31, 34, 60, 62], we leverage the labels to establish a curriculum sequence and mitigate the bias for pseudo labels towards camera labels.

**Curriculum learning.** The seminal work of [2] introduces a curriculum learning strategy that trains a model using easy examples in early stages and with hard ones in later stages. Since then, the curriculum learning paradigm [40, 43] is adopted for various applications, including object detection [39], semantic segmentation [36], and image synthesis [23]. The main difference between these methods lies in how they define easy and hard examples. Previous methods typically define specific criteria to establish curriculum sequences, *e.g.*, by measuring distances to object boundaries for semantic segmentation [30], computing loss values of training samples [17], and employing a module for estimating difficulties of samples [47]. More specific for UDA segmentation, a domain discriminator [36] and a pixel-wise label distribution [50] are used to define easy and hard samples. On the other hand, we incorporate camera labels of person images to facilitate a curriculum learning paradigm for UDA reID. We conjecture that camera topologies play a significant role in learning discriminative features, particularly for the task of person reID. To our knowledge, no previous approaches have incorporated camera topologies to set a curriculum.

## 3. Method

### 3.1. Overview

We provide in Fig. 2 an overview of our framework for UDA reID. We first divide a target dataset into multiple subsets by leveraging camera labels of person images. The camera-driven scheduler takes the target subsets, along with source images as inputs, to establish a curriculum sequence. Within each curriculum sequence, we adopt a self-training scheme [10] and alternate between clustering and fine-tuning. Specifically, we apply a clustering algorithm on person features extracted from target images to generate pseudo ID labels, and further fine-tune our model using a joint set of source and target images [5, 11, 13, 19, 21, 31, 51, 52, 53, 57]. We incorporate

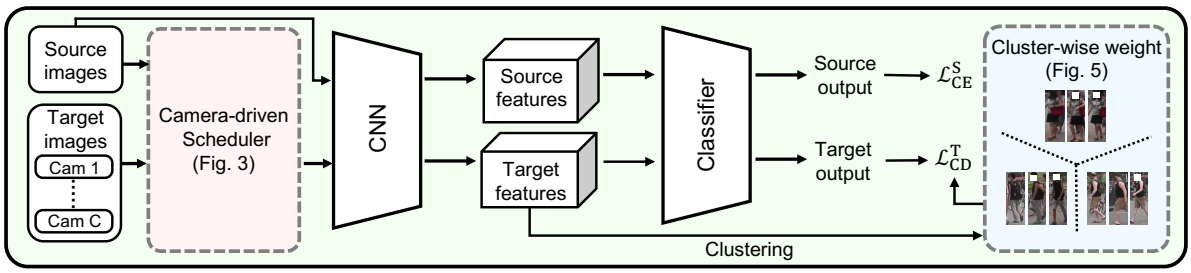


Figure 2: An overview of our framework. We divide target images into multiple subsets based on camera labels. The camera-driven scheduler takes the subsets of the target domain, along with source images as inputs, to establish a curriculum sequence. We train our model progressively with CD loss for a target domain $\mathcal{L}_{\text{CD}}^{\text{T}}$, along with the cross-entropy term [58] for a source domain $\mathcal{L}_{\text{CE}}^{\text{S}}$. See text for more details.

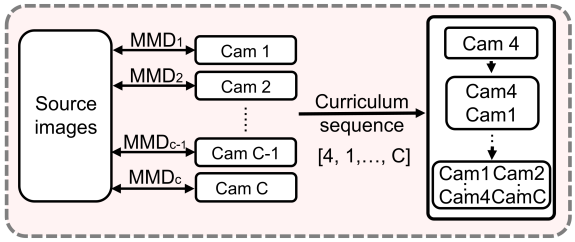


Figure 3: An illumination of a camera-driven scheduler. We compute pairwise MMDs between source domain and all target subsets, establishing a curriculum sequence. We initially train our model with a single subset and gradually expand a training set by adding subsets in the sequence.

the CD loss for target images to consider the diversity of camera labels within each cluster. At test time, we compute L2 distances between query and gallery person representations to perform cross-camera matching. Note that the camera labels are used during training only.

### 3.2. CaCL

Given a target dataset, obtained from $C$ different cameras, we first divide the target dataset into multiple subsets by exploiting camera labels of person images. Concretely, we denote by $\mathcal{D}^{\text{S}}$ and $\mathcal{D}^{\text{T}}$ sets of images in the source and target domain datasets, respectively. We divide $\mathcal{D}^{\text{T}}$ into total $C$ number of non-overlapping subsets w.r.t camera labels, where each subset is denoted by $\mathcal{D}_c^{\text{T}}$, and $\mathcal{D}^{\text{T}} = \mathcal{D}_1^{\text{T}} \cup \mathcal{D}_2^{\text{T}} \cup \cdots \cup \mathcal{D}_C^{\text{T}}$. We start training a model using a single subset in the first curriculum stage, and incrementally expand the training set to $c$ subsets in the $c$-th stage for $c = 1, \ldots, C$, according to a curriculum sequence, computed by a camera-driven scheduling rule. That is, the curriculum sequence determines which subsets are used to increment the training set at each stage (Fig. 3). At each curriculum stage, we employ a self-training scheme [10] that alternates between clustering and fine-tuning steps.

**Camera-driven scheduler (Fig. 3).** Setting an effective training sequence plays an important role in curriculum learning [40, 43]. In the context of our approach to leveraging camera labels, the scheduling is equivalent to determining which camera in the target domain is easier to learn for a reID model trained on the source domain. We assume that knowledge transfer between domains of similar distributions is typically easier than the opposite case. We implement this idea using the MMD [16] that computes distributional discrepancies between different domains. Specifically, we compute pairwise MMDs between the source dataset, $\mathcal{D}^{\text{S}}$, and target subsets, $\mathcal{D}_c^{\text{T}}$, by mapping the samples to the reproducing kernel Hilbert space $\mathcal{H}$ with a function $\phi(\cdot)$ associated with Gaussian kernel, as follows:

$$\text{MMD}_c = \left\| \frac{1}{|\mathcal{D}^{\text{S}}|} \sum_{\mathbf{x}_i^{\text{S}} \in \mathcal{D}^{\text{S}}} \phi(\mathbf{x}_i^{\text{S}}) - \frac{1}{|\mathcal{D}_c^{\text{T}}|} \sum_{\mathbf{x}_j^{\text{T}} \in \mathcal{D}_c^{T}} \phi(\mathbf{x}_j^{\text{T}}) \right\|_{\mathcal{H}}^2, \tag{1}$$

where $\mathbf{x}_i^{\text{S}}$ and $\mathbf{x}_j^{\text{T}}$ denote the $i$-th and $j$-th sample in $\mathcal{D}^{\text{S}}$ and $\mathcal{D}_c^{\text{T}}$, respectively, and $|\cdot|$ counts the total number of samples within a set. We establish a curriculum sequence by sorting the pairwise MMDs in an ascending order, that is, the closer subset w.r.t the source domain in terms of MMD is exploited earlier.

CaCL with a camera-driven scheduler provides the following advantages for UDA reID: First, it allows a smooth adaptation from source to target domains in a progressive manner. CaCL leverages a subset within a target domain that depicts a similar distribution with the source domain in earlier training stages, facilitating a smooth adaptation, compared to previous approaches using domain-level regimes. Second, our model starts to learn from person images obtained using a single camera, typically showing a weaker extent of intra-class variations, then progressively expands to other images captured from multiple cameras. This gradual expansion from simple to diverse scenarios encourages our model to better handle the inter-camera variations.

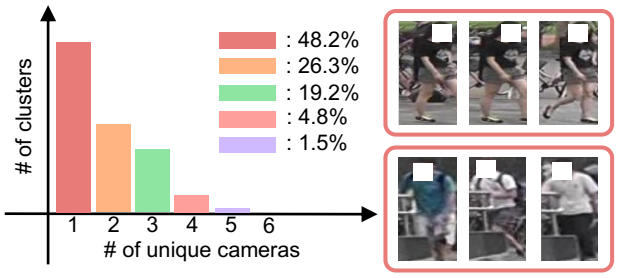


Figure 4: **Left**: Distribution of the number of clusters, across the number of unique cameras. We obtain the result on Market1501 [55] using a reID model trained on MSMT17 [45]. Since the reID model trained on a single domain fails to generalize on other domains, most clusters simply contain person images captured by a single camera. **Right**: Examples of person images within the same cluster. We can see that the person images do not show diverse intra-class variations (top), and the clustering results are easily influenced by distracting cues (*e.g.*, occlusion in the left).

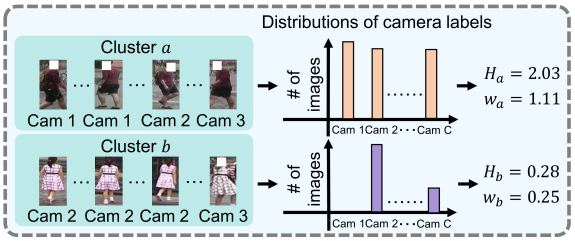


Figure 5: An illustration of a detailed procedure for computing cluster-wise weighting factors. We compute the entropy of a camera distribution for each cluster, and then assign large weights for clusters with high entropy values. See text for details.

### 3.3. CD loss

To generate pseudo ID labels for unlabeled target domain, previous methods [5, 11, 13, 19, 21, 31, 51, 52, 53, 57] apply a clustering algorithm on person representations of target images. However, we have observed that clustering results for target images are highly biased toward camera labels. We show in Fig. 4 that most clusters (up to 48.2%) are trivial ones that contain images obtained using a single camera. Directly training with the biased pseudo ID labels using standard cross-entropy [58] and triplet [20] losses might be suboptimal, since the trivial clusters can dominate the adaptation process. Note that this is particularly important for reID that performs cross-camera image retrieval. To alleviate this issue, we propose to discard the trivial clusters, while encouraging the clusters of images from various cameras to involve more in feature learning. To this end, we measure the entropy of a camera distribution for each cluster, as follows:

**Algorithm 1** Training

**Require:** $N_c$: the number of iterations at $c$-th stage; $M_c$: an interval of generating pseudo labels at $c$-th stage; $\mathcal{A}$: an empty set; $C$: the number of cameras.
**Input:** Source dataset $\mathcal{D}^{\mathrm{S}}$; Target subsets $\mathcal{D}_1^{\mathrm{T}}, \mathcal{D}_2^{\mathrm{T}}, \ldots, \mathcal{D}_C^{\mathrm{T}}$.
**Output:** A trained reID network.

```
Pre-train a network using D^S.
Compute MMD between D^S and D_c^T [Eq. (1)].
Determine a curriculum sequence of target subsets and obtain a list of ordered subsets O
for c = 1 to C do
    A ← A ∪ O(c)
    while i ≤ N_c do
        if (i mod M_c) = 0 then
            Cluster images of A and generate pseudo labels.
            Measure cluster-wise camera entropy H_l [Eq. (2)].
            Obtain cluster-wise weighting factor w_l [Eq. (3)].
        end if
        Sample a mini-batch from D^S and A.
        if c = 1 then
            Update the network using L without w_l [Eq. (5)].
        else
            Update the network using L with w_l [Eq. (5)].
        end if
    end while
end for
```

$$H_l = -\sum_c r_l(c) \log(r_l(c)), \tag{2}$$

where $r_l(c) = \frac{n_l(c)}{\sum_c n_l(c)}$, and $n_l(c)$ is the number of images captured by the $c$-th camera within the $l$-th cluster. With the entropy for each cluster, we define cluster-wise weighting factor:

$$w_l = \log(H_l + 1). \tag{3}$$

Namely, for clusters of images obtained from the same camera (*i.e.*, $H_l = 0$), the weighting factor becomes zero and discards the clusters for training. On the other hand, for clusters of images captured by different cameras, the weighting factor encourages images in the clusters to involve more for feature learning (Fig. 5).

We incorporate cluster-wise weights, $w_l$, to enhance cross-entropy and triplet terms used for training reID models. Concretely, given a person image $\mathbf{x}_i^{\mathrm{T}}$ assigned to the $l$-th cluster, we define the CD cross-entropy term as follows:

$$\mathcal{L}_{\mathrm{CDC}}^{\mathrm{T}} = \mathbb{E}[-w_l \log(p(l|\mathbf{x}_i^{\mathrm{T}}))], \tag{4}$$

where $p(l|\mathbf{x}_i^{\mathrm{T}})$ is a softmax probability of $\mathbf{x}_i^{\mathrm{T}}$ being classified to the $l$-th pseudo ID label. The CD triplet term $\mathcal{L}_{\mathrm{CDT}}^{\mathrm{T}}$ is defined similarly. Note that for the first curriculum sequence, where input target images are taken from a single camera, we omit the weighting factors and employ vanilla cross-entropy and triplet losses for training.

Table 1: Quantitative comparisons with the state of the art on a real-to-real scenario. Numbers in bold indicate the best performance and underscored ones indicate the second best. Results in parentheses are obtained with the source codes provided by the authors.

| Methods | Reference | MSMT17-to-Market1501 | | | | Market1501-to-MSMT17 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R1 | R5 | R10 | mAP | R1 | R5 | R10 |
| MMT [11] | ICLR 2020 | 75.6 | 89.3 | 95.8 | 97.5 | 22.9 | 49.2 | 63.1 | 68.8 |
| SpCL [13] | NeurIPS 2020 | 77.5 | 89.7 | 96.1 | 97.6 | 26.8 | 53.7 | 65.0 | 69.8 |
| UNRN [52] | AAAI 2021 | (78.3) | (90.4) | (96.5) | (97.9) | 25.3 | 52.4 | 64.7 | 69.7 |
| GLT [53] | CVPR 2021 | (79.3) | (90.7) | (96.5) | (98.0) | 26.5 | 56.6 | 67.5 | 72.0 |
| HCD [57] | ICCV 2021 | 80.2 | 91.4 | - | - | 28.4 | 54.9 | - | - |
| IDM [5] | ICCV 2021 | 82.1 | 92.4 | 97.5 | 98.4 | 33.5 | 61.3 | 73.9 | 78.4 |
| RESL [31] | AAAI 2022 | - | - | - | - | 33.6 | 64.8 | 74.6 | 79.6 |
| Ours | | **84.7** | **93.8** | **97.7** | **98.6** | **36.5** | **66.6** | **75.3** | **80.1** |

Table 2: Quantitative comparisons with the state of the art on a synthetic-to-real scenario. Numbers in bold indicate the best performance and underscored ones indicate the second best.

| Methods | Reference | PersonX-to-Market1501 | | | | PersonX-to-MSMT17 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R1 | R5 | R10 | mAP | R1 | R5 | R10 |
| MMT [11] | ICLR 2020 | 71.0 | 86.5 | 94.8 | 97.0 | 17.7 | 39.1 | 52.6 | 58.5 |
| SpCL [13] | NeurIPS 2020 | 73.8 | 88.0 | 95.3 | 96.9 | 22.7 | 47.7 | 60.0 | 65.5 |
| IDM [5] | ICCV 2021 | 81.3 | 92.0 | 97.4 | 98.2 | 30.3 | 58.4 | 70.7 | 75.5 |
| Ours | | **82.3** | **92.8** | **97.6** | **98.6** | **36.2** | **66.9** | **69.4** | **80.9** |
| Methods | Reference | Unreal-to-Market1501 | | | | Unreal-to-MSMT17 | | | |
| | | mAP | R1 | R5 | R10 | mAP | R1 | R5 | R10 |
| JVTC [27] | ECCV 2020 | 78.3 | 90.8 | - | - | 25.0 | 53.7 | - | - |
| IDM [5] | ICCV 2021 | 83.2 | 92.8 | 97.3 | 98.2 | 38.3 | 67.3 | 78.4 | 82.6 |
| Ours | | **84.0** | **93.3** | **97.6** | **98.5** | **40.3** | **70.0** | **80.5** | **84.0** |

### 3.4. Overall training

We pre-train a reID model using ground-truth ID labels of source images using conventional cross-entropy [58] and triplet [20] losses. We establish a curriculum with a camera-driven scheduler, and then perform clustering to obtain pseudo ID labels for target images. During fine-tuning, we exploit both source and target domains jointly, following [5, 11, 21, 52, 53, 57]. We adopt the cross-entropy loss ($\mathcal{L}_{\text{CE}}^{\text{S}}$) for source images, and the CD term ($\mathcal{L}_{\text{CD}}^{\text{T}}$) for target images, where the CD loss consists of CD cross-entropy and CD triplet terms. At each fine-tuning stage, we optimize a reID network with the overall objective as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}^{\text{S}} + \mathcal{L}_{\text{CD}}^{\text{T}}. \tag{5}$$

We summarize in Algorithm 1 an overall training process of our approach.

## 4. Experiments

### 4.1. Implementation details

**Dataset and evaluation metric.** We use four person reID datasets in our experiments, including Market1501 [55], MSMT17 [45], PersonX [41] and Unreal [49], where PersonX and Unreal provide synthetic images and corresponding ID labels. Market1501 contains pedestrian images of 1,501 IDs, captured by 6 cameras, where it consists of 12,936 images of 751 IDs for training and 19,732 images of 750 IDs for testing. MSMT17 contains 126,441 images, obtained from 15 cameras, where it consists of 32,621 images of 1,041 IDs and 93,820 images of 6,120 IDs for training and testing, respectively. PersonX and Unreal provide 9,840 and 130,244 images, respectively, for training. Following the evaluation protocol in UDA reID [5, 13, 31, 52, 53], we apply our approach to real-to-real and synthetic-to-real scenarios. We report the mean average precision (mAP) and cumulative matching characteristics (CMC) at rank-1, rank-5, and rank-10 for evaluation.

**Training.** We adopt ResNet-50 [18], pre-trained for ImageNet classification [6], as a backbone network, where we use domain-specific BNs [4] following [5, 52, 53]. We train ResNet-50 with a source dataset, and use it as an initial reID model for UDA reID. We train the model for 4 epochs for each curriculum stage, except for the final stage, where we use 30 epochs, with the learning rate of $3.5 \times 10^{-4}$. Following [5, 52, 53], we set the batch size to 128, with 64 images from each domain. We use the Adam optimizer [24] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and employ the XBM tech-

Table 3: Quantitative comparisons of variants of our model on Market1501 [55]-to-MSMT17 [45] and MSMT17-to-Market1501. Numbers in bold indicate the best performance and underscored ones indicate the second best. M: Market1501, MS: MSMT17, RS: Random sequence, CTL: Cross-entropy [58] and triplet [20] losses, CDL: CD loss.

| Curriculum | | Loss | | M-to-MS | | MS-to-M | |
|---|---|---|---|---|---|---|---|
| RS | CaCL | CTL | CDL | mAP | rank-1 | mAP | rank-1 |
| | | ✓ | | 23.4 | 50.4 | 79.2 | 92.3 |
| ✓ | | ✓ | | 24.4 | 51.3 | 79.4 | 92.4 |
| ✓ | | | ✓ | 30.4 | 58.5 | 81.4 | 92.8 |
| | ✓ | ✓ | | 31.1 | 59.9 | 81.2 | 92.7 |
| | | | ✓ | 30.1 | 58.3 | 81.8 | 93.0 |
| | ✓ | | ✓ | **36.5** | **66.6** | **84.7** | **93.8** |

Table 4: Quantitative comparisons between CD loss and UGID [52] loss on real-to-real and synthetic-to-real scenarios. Numbers in bold indicate the best performance and underscored ones indicate the second best. M: Market1501, MS: MSMT17, PX: PersonX, U: Unreal, CDL: Camera-diversity loss.

| Variants | M-to-MS | | MS-to-M | |
|---|---|---|---|---|
| | mAP | rank-1 | mAP | rank-1 |
| UGID [52] | 32.3 | 61.0 | 82.2 | 92.8 |
| CDL | 36.5 | 66.6 | 84.7 | 93.8 |
| CDL+UGID [52] | **38.4** | **67.9** | **85.2** | **94.1** |
| Variants | PX-to-M | | U-to-MS | |
| | mAP | rank-1 | mAP | rank-1 |
| UGID [52] | 81.2 | 91.9 | 35.2 | 63.4 |
| CDL | 82.3 | 92.8 | 40.3 | 70.0 |
| CDL+UGID [52] | **83.1** | **93.2** | **41.2** | **71.4** |

nique [44] for triplet losses throughout all experiments, as done in [5]. We use the DBSCAN [9] algorithm to cluster target images and generate pseudo ID labels, where we update the pseudo labels at every 3 epochs. Following [5, 52, 53], we resize the person image to the size of $256 \times 128$ and apply data augmentation techniques, including random flipping, random cropping, and random erasing [59]. Detailed descriptions for hyperparameter settings are available in the supplement.

## 4.2. Comparison with the state of the art

We compare our method with the state of the art on the real-to-real scenario in Table 1. Overall, we can see from the results that our approach outperforms other methods on all benchmarks. UNRN [52] focuses on leveraging reliable labels, but does not consider the camera bias problem of pseudo labels. In contrast to UNRN, ours addresses the camera bias in pseudo labels, outperforming UNRN in all benchmarks by significant margins. IDM [5] generates intermediate domains and leverages them to bridge source and target domains. However, it exploits all target images jointly during the adaptation process. In contrast to IDM, we address the large distribution gap of camera topologies between domains, by using camera labels of target images, outperforming IDM on all benchmarks. RESL [31] also exploits camera labels of target images to train translation networks [61] that generate multiple images of the same person in the style of different cameras. On the contrary, we leverage camera labels of target images to establish a curriculum sequence and address the camera bias of pseudo labels. Our method outperforms RESL even without using the translation networks, indicating that our framework effectively leverages the camera labels to perform UDA reID.

We provide in Table 2 a quantitative comparison between ours and state-of-art methods in the synthetic-to-real scenario. The results demonstrate that ours can effectively transfer the knowledge learned from the source domain to the target one, even for the synthetic-to-real scenario.

## 4.3. Discussion

**Ablation study.** We present in Table 3 an ablation analysis for each component of our method on Market1501-to-MSMT17 and MSMT17-to-Market1501. We report mAP and rank-1 scores for variants of our model. To validate the effectiveness of a camera-driven scheduler, we also provide results of setting a curriculum sequence randomly (RS), and report the scores averaged over 5 trials. For the variants trained using cross-entropy and triplet losses (CTL), we exclude the weighting factor within the CD loss for target images. We can see from the first and second rows that establishing a curriculum in a random sequence boosts the performance marginally, as this does not consider the discrepancies between target subsets and the source dataset. By exploiting the camera-driven scheduler in the fourth row, we boost the performance drastically. This coincides with findings reported in [40, 43], that the result of incorporating curriculum depends on how the curriculum is designed. In this context, camera-driven scheduling rule provides a beneficial sequence for UDA reID, and enhances the adaptation performance. By comparing the fourth and the sixth rows, we can see that the CD loss further enhances the performance, confirming that incorporating cluster-wise weighting factors to selectively involve clusters is effective for adaptation. We can see from the first and fifth rows that CD loss still performs better than CTL even without a CaCL.

**Camera-diversity loss.** Similar to the weighting scheme in the CD loss, the work of [52] employs an UGID loss to re-weight the loss terms by measuring uncertainty among pseudo ID labels, and assigns large weights to the labels with low uncertainty values. For a comparison with the weighting scheme proposed in [52], we show in Table 4 results of models trained using the CD loss, the UGID-

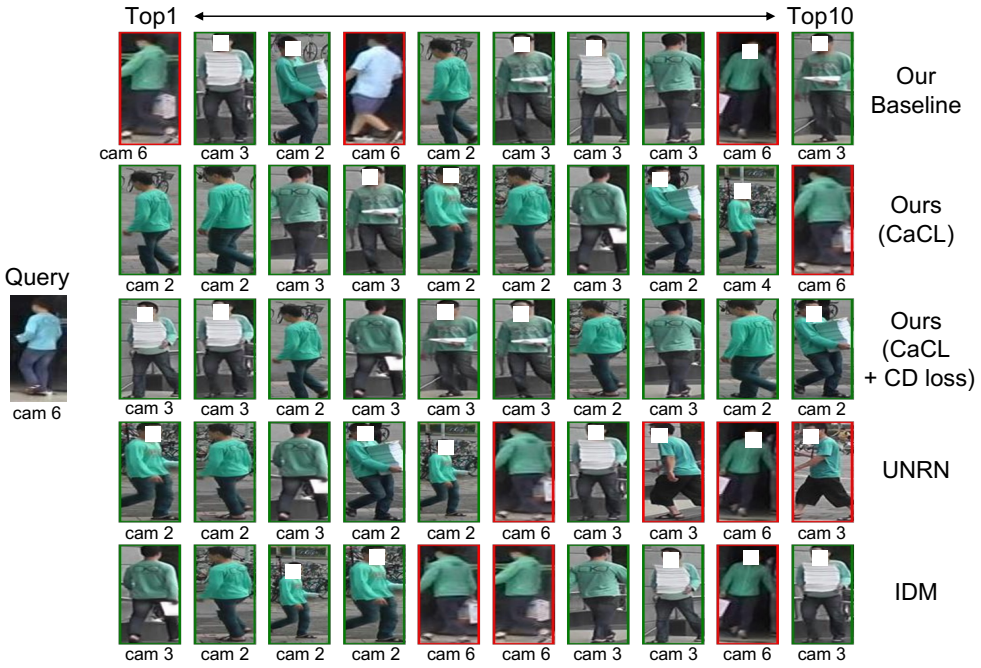


Figure 6: Visual comparisons of retrieval results on MSMT17 [45]-to-Market1501 [55]. Results with green boxes have the same identity as the query, while those with red boxes do not. (Best viewed in color.)

weighted loss, and a combination of the two. We simply multiply corresponding weight values from CD and UGID terms to exploit both losses. By comparing the first and the second rows, we can see that the model trained using the CD loss performs better than the one trained with the UGID for both cases. This suggests that addressing the inter-camera variations is more effective for UDA reID compared to handling samples with reliable labels. Note that computing uncertainty values for the UGID loss requires multiple reID models to measure prediction consistency, and thus demands additional computational complexity compared to the CD loss. We can also see that exploiting both losses in the third row shows the best performance, because the two weighting factors can complement each other. This suggests that combining our framework with other methods could lead to significant performance improvements in UDA reID.

**Qualitative analysis.** We show in Fig. 6 visual comparisons of retrieval results with the state of the art and variants of our model on MSMT17 [45]-to-Market1501 [55]. The baseline is trained using the vanilla cross-entropy [58] and triplet [20] losses without exploiting CaCL and the CD loss. We can see from the first and second rows that CaCL is more effective to retrieve person images, compared to the baseline. The third row shows that CaCL with the CD loss retrieves person images with the same IDs as the query correctly without the camera bias problem, confirming the effectiveness of the CD loss. Last three rows compare ours with other approaches (UNRN [52] and IDM [5]). We can observe that they also retrieve person images with different IDs as the query, and suffer from the camera bias problem. In contrast, ours obtains accurate retrieval results, suggesting that it is robust to the camera bias problem of pseudo labels, effectively reducing inter-camera variations.

## 5. Conclusion

We have presented a novel approach for UDA reID that performs a progressive adaptation by leveraging camera labels of person images. We propose a CaCL framework, gradually transferring the knowledge learned from a source domain to a target one, while addressing the large distribution gap of camera topologies between domains. We have also introduced a novel CD loss, mitigating a camera bias in pseudo labels and handling inter-camera variations, while progressively adapting a reID model from source to target domains. Experimental results show the effectiveness of our framework, setting a new state of the art on standard benchmarks.

**Acknowledgments.** This work was partly supported by the IITP and NRF grants funded by the Korea government(MSIT) (No.RS-2022-00143524, Development of Fundamental Technology and Integrated Solution for Next-Generation Automatic Artificial Intelligence System, No. 2023R1A2C2004306), and the Yonsei Signature Research Cluster Program of 2023 (2023-22-0008).

## References


[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 3

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009. 3

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[4] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, 2019. 6

[5] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. IDM: An intermediate domain module for domain adaptive person re-id. In *ICCV*, 2021. 2, 3, 5, 6, 7, 8

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[7] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018. 2

[8] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACM MM*, 2014. 3

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 2, 7

[10] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: clustering and fine-tuning. *TOMM*, 2018. 3, 4

[11] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 2, 3, 5, 6

[12] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. In *NeurIPS*, 2018. 3

[13] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020. 2, 3, 5, 6

[14] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NeurIPS*, 2014. 2

[16] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *NeurIPS*, 2006. 2, 4

[17] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[19] Tao He, Leqi Shen, Yuchen Guo, Guiguang Ding, and Zhenhua Guo. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In *AAAI*, 2022. 2, 3, 5

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 5, 6, 7, 8

[21] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. Towards discriminative representation learning for unsupervised person re-identification. In *ICCV*, 2021. 2, 3, 5, 6

[22] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 3

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 3

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[25] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*, 2017. 3

[26] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 3

[27] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *ECCV*, 2020. 6

[28] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *TPAMI*, 2020. 3

[29] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2

[30] Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *CVPR*, 2017. 3

[31] Zongyi Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Qian Wang, and Fengfan Zhou. Reliability exploration with self-ensemble learning for domain adaptive person re-identification. In *AAAI*, 2022. 2, 3, 5, 6, 7

[32] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *PR*, 2019. 3

[33] Andy J Ma, Pong C Yuen, and Jiawei Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *CVPR*, 2013. 2

[34] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *ICCV*, 2019. 3

[35] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop*,

2016. 2

[36] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019. 3

[37] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 3

[38] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPR Workshop*, 2017. 3

[39] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 3

[40] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *IJCV*, 2022. 3, 4, 7

[41] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, 2019. 2, 6

[42] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold Mixup: Better representations by interpolating hidden states. In *ICML*, 2019. 3

[43] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *TPAMI*, 2021. 3, 4, 7

[44] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *CVPR*, 2020. 7

[45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 2, 5, 6, 7, 8

[46] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 3

[47] Luyu Yang, Yogesh Balaji, Ser-Nam Lim, and Abhinav Shrivastava. Curriculum manager for source selection in multi-source domain adaptation. In *ECCV*, 2020. 3

[48] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 2021. 1

[49] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In *CVPR*, 2021. 2, 6

[50] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017. 3

[51] Fang Zhao, Shengcai Liao, Guo-Sen Xie, Jian Zhao, Kaihao Zhang, and Ling Shao. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *ECCV*, 2020. 2, 3, 5

[52] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zheng-Jun Zha. Exploiting sample uncertainty for domain adaptive person re-identification. In *AAAI*, 2021. 2, 3, 5, 6, 7, 8

[53] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *CVPR*, 2021. 2, 3, 5, 6, 7

[54] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *TIP*, 2019. 3

[55] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2, 5, 6, 7, 8

[56] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[57] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *ICCV*, 2021. 2, 3, 5, 6

[58] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *TOMM*, 2017. 2, 4, 5, 6, 7, 8

[59] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 7

[60] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 3

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 7

[62] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *ECCV*, 2020. 3