

Joint Noise-Tolerant Learning and Meta Camera Shift Adaptation for Unsupervised Person Re-Identification

Fengxiang Yang^{1*}, Zhun Zhong², Zhiming Luo^{1†}, Yuanzheng Cai⁴, Yaojin Lin⁵, Shaozi Li^{1,3†}, Nicu Sebe²

¹ Department of Artificial Intelligence, School of Informatics, Xiamen University

² Department of Information Engineering and Computer Science, University of Trento

³ Institute of Artificial Intelligence, Xiamen University

⁴ Minjiang University ⁵ Minnan Normal University

Project: https://github.com/FlyngRoastDuck/MetaCam_DSCE

Abstract

This paper considers the problem of unsupervised person re-identification (re-ID), which aims to learn discriminative models with unlabeled data. One popular method is to obtain pseudo-label by clustering and use them to optimize the model. Although this kind of approach has shown promising accuracy, it is hampered by 1) noisy labels produced by clustering and 2) feature variations caused by camera shift. The former will lead to incorrect optimization and thus hinders the model accuracy. The latter will result in assigning the intra-class samples of different cameras to different pseudo-label, making the model sensitive to camera variations. In this paper, we propose a unified framework to solve both problems. Concretely, we propose a Dynamic and Symmetric Cross-Entropy loss (DSCE) to deal with noisy samples and a camera-aware meta-learning algorithm (MetaCam) to adapt camera shift. DSCE can alleviate the negative effects of noisy samples and accommodate the change of clusters after each clustering step. MetaCam simulates cross-camera constraint by splitting the training data into meta-train and meta-test based on camera IDs. With the interacted gradient from meta-train and meta-test, the model is enforced to learn camera-invariant features. Extensive experiments on three re-ID benchmarks show the effectiveness and the complementarity of the proposed DSCE and MetaCam. Our method outperforms the state-of-the-art methods on both fully unsupervised re-ID and unsupervised domain adaptive re-ID.

1. Introduction

Person re-identification (re-ID) attempts to find matched pedestrians of a query in a non-overlapping camera sys-

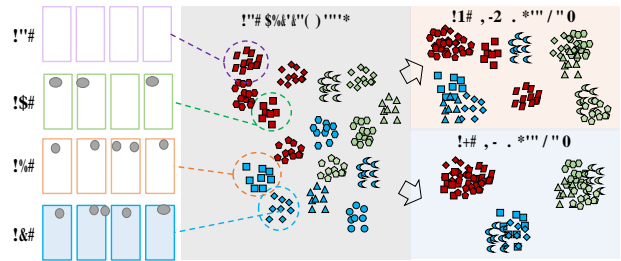


Figure 1. Illustration of camera variations in person re-ID (a) and the comparison between methods trained without or with the proposed MetaCam (b) and (c), respectively). Different colors represent different identities and different shapes indicate different camera IDs. At the initial state, samples under different cameras may suffer from appearance changes of viewpoints ((1) & (2)), illumination ((3) & (4)), and other factors. Without considering this factor, the trained model may be sensitive to camera variations and may wrongly split intra-class features to different centers. Our proposed MetaCam enables the model to learn camera-invariant features by explicitly considering cross-camera constraint.

tem. Recent CNN-based works [31, 35] have achieved impressive accuracies, but their success is largely dependent on sufficient annotated data that require a lot of labeling cost. In contrast, it is relatively easy to obtain a large collection of unlabeled person images, fostering the study of unsupervised re-ID. Commonly, unsupervised re-ID can be divided into two categories depending on whether using an extra labeled data, *i.e.*, unsupervised domain adaptation (UDA) [37, 49, 7] and fully unsupervised re-ID (FU) [19, 20, 42]. In UDA, we are given a labeled source domain and an unlabeled target domain. The data of two domains have different distributions and are used to train a model that generalizes well on the target domain. The fully unsupervised re-ID is more challenging since only unlabeled images are provided to train a deep model. In this study, we will mainly focus on this setting, and call it as unsupervised re-ID for simplicity.

*Equal contribution: yangfx@stu.xmu.edu.cn

†Corresponding author: {zhiming.luo, szlig}@xmu.edu.cn

Recent popular unsupervised re-ID methods [19, 34, 20, 42] mainly adopt clustering to generate pseudo-label for unlabeled samples, enabling the training of the model in a supervised manner. Pseudo-label generation and model training are applied iteratively to train an accurate deep model. Despite their effectiveness, existing methods often ignore two important factors during this process. (1) *Noisy labels brought by clustering*. The clustering algorithm cannot ensure intra-samples to be assigned with the same identity, which inevitably will introduce noisy labels in the labeling step. The errors of noisy labels will be accumulated during training, thereby hindering the model accuracy. (2) *Feature variations caused by camera shifts*. As shown in Fig. 1, intra-class samples under different cameras may suffer from the changes of viewpoint (e.g., (1) and (2) in Fig. 1), illumination (e.g., (3) and (4) in Fig. 1), and other environmental factors. At the start of unsupervised learning (“initial state” in Fig. 1), these significant variations will cause large gaps between the intra-class features of different cameras. In such a situation, it is difficult for the clustering algorithm to cluster samples with the same identity from all cameras into the same cluster. Consequently, training with the samples mined by the clustering will lead to unexpected separation for intra-class samples (“w/o MetaCam” in Fig. 1) and the model might be sensitive to camera variations during testing. In this paper, we attempt to solve the above two crucial problems for robust unsupervised re-ID.

For the first issue, we try to adopt the technique of learning with noisy labels (LNL) for robust training. LNL is well-studied in image classification, however, most of the existing methods cannot be directly applied to our scenario. This is because the centers and pseudo-label will change after each clustering step. To overcome this difficulty, this paper proposes a *dynamic and symmetric cross-entropy loss* (DSCE) for unsupervised re-ID. We maintain a feature memory to store all image features, which enables us to dynamically build new class centers and thus to be adaptable to the change of clusters. With the dynamic centers, a robust loss function is proposed for mitigating the negative effects caused by noisy samples.

For the second issue, we attempt to explicitly consider camera-invariant constraint during training. Indeed, person re-ID is a cross-camera retrieval process, aiming to learn a model that can well discriminate samples under different cameras. If a model trained with samples from some of the cameras can also generalize to distinguish samples from the rest of the cameras, then, we could obtain a model that can extract the intrinsic feature without camera-specific bias and is robust to camera changes. Inspired by this, this paper introduces a camera-aware meta-learning (MetaCam), which aims to learn camera-invariant representations by simulating the cross-camera re-identification process during training. Specifically, MetaCam separates the training data into

meta-train and meta-test, ensuring that they belong to entirely different cameras. We then enforce the model to learn camera-invariant features under both camera settings by updating the model with meta-train and validating the updated model with meta-test. Along with learning from different meta divisions, the model is gradually optimized to generalize well under all cameras. As shown in Fig. 1, MetaCam gathers intra-class features of different cameras into the same cluster, which is beneficial for mining pseudo-label and learning camera-invariant features. In summary, our main contributions can be summarized in three aspects:

- We propose a dynamic and symmetric loss (DSCE), which enables the model to be robust to noisy labels during training in the context of changes of clusters and thus promotes the model performance.
- We propose a camera-aware meta-learning algorithm (MetaCam) for adapting the shifts caused by cameras. By simulating the cross-camera searching process during training, MetaCam can effectively improve the robustness of the model to camera variations.
- We introduce a unified framework that can jointly take advantage of the proposed DSCE and MetaCam, enabling us to learn a more robust re-ID model.

Extensive experiments on three large-scale datasets demonstrate the advantages of our DSCE and MetaCam for the fully unsupervised re-ID. Besides, further experiments on the UDA setting show that our method can also achieve state of the art.

2. Related Work

2.1. Unsupervised Person Re-ID

Unsupervised person re-ID can be categorized into *Fully Unsupervised Re-ID* (FU) [19, 20, 42] and *Unsupervised Domain Adaptation* (UDA) [49, 41, 18, 24]. The former tries to train a re-ID model with only unlabeled data while the latter attempts to leverage labeled source data and unlabeled target data to train a model for the target domain. Although training under different data conditions, most methods for UDA and FU adopt similar learning strategies, which can be summarized into methods based on pseudo-label discovery [6, 7, 19, 42] and methods based on alignment [2, 37, 24, 52]. The methods based on pseudo-label discovery rely on the iteration of pseudo-label mining and model fine-tuning, such as BUC [19], SSG [6], HCT [42] and SpCL [8]. Despite their success, these methods might suffer from noisy samples obtained in the pseudo-label mining process. The alignment-based methods try to align distribution shift (e.g., camera shift or domain shift) in image level or feature level. In FU, we only have one dataset, therefore we mainly focus on reducing the camera shift in dataset. Zhong *et al.* [48] propose to align camera shift of

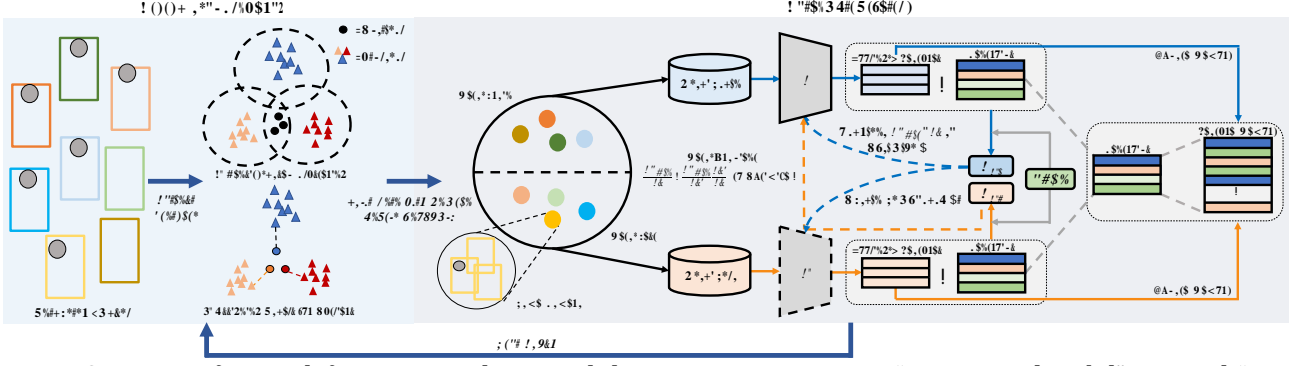


Figure 2. Our training framework for unsupervised re-ID includes two training stages, *i.e.*, “Mining Pseudo-Label” stage and “Meta-Optimization”. The first stage assigns samples with pseudo-label based on DBSCAN [4]. The second stage splits the training data into meta-train and meta-test sets based on camera labels and optimizes the model with the proposed meta-learning strategy. This camera-aware meta-learning (MetaCam) encourages the model to learn camera-invariant features. To reduce the negative impact of noisy labels, we also propose a dynamic and symmetric cross-entropy loss (DSCE) that is used for both meta-train and meta-test data.

the target samples by camera-to-camera style transfer [51]. Yu *et al.* [41] attempt to align the camera shift with 2-Wasserstein distance. The former is sensitive to distorted images in the generation process while the latter should be implemented under the precondition that features of pedestrians satisfy Gaussian distribution. Different from these two works, this paper aligns the camera shift by simulating the cross-camera process during training, instead of generating virtual images or learning based on prior assumptions.

2.2. Learning with Noisy Labels

Recent studies on learning with noisy labels (LNL) can mainly be categorized into sample re-weighting methods [28, 1, 11, 40], label correction methods [15, 33, 39] and robust loss designing methods [9, 36, 43]. Re-weighting methods assign clean samples with higher weights during the training process. Shu *et al.* [28] attempt to train an additional network with meta-learning for sample re-weighting. However, it requires additional clean data to be meta-test set during optimization, which may not be available in real-world applications. Co-teaching [11] proposes to remove noisy samples with a pair of networks. It can also be regarded as re-weighting methods, since the removed samples can be regarded as being assigned with zero weights. Label correction methods [15, 33, 39] try to identify noisy samples and correct their labels during training. These methods also require additional clean data to assist the model to correct labels. For robust loss functions, Ghosh *et al.* [9] propose a theory to check the robustness of loss functions and find that mean absolute error loss (MAE) is robust to noisy labels. However, the gradient saturation of MAE prevents it from obtaining satisfactory performance. To address this problem, Wang *et al.* [36] propose a symmetric cross-entropy loss for robust learning. Our work leverages the finding of [9, 36] and proposes a dynamic and symmetric cross-entropy loss to resist noisy labels produced by clustering.

2.3. Meta Learning

Meta-learning aims to learn new tasks with limited training samples and can be classified into optimizing-based [5, 23, 25], model-based [27, 22] and metric-based [31, 29, 32] methods. Our work is closely related to optimizing-based meta-learning methods, which try to obtain a good initialized weight for fast adaptation for new tasks. Finn *et al.* [5] propose model-agnostic meta-learning (MAML) to acquire the ideal weight by stimulating the learning process of new tasks with meta-test set. Subsequently, Reptile [23] speeds up the learning process of MAML with a first-order approximation. Recently, MAML has been widely used in computer vision tasks such as noisy label learning [28], domain generalization [16] and face recognition [10]. For the line of unsupervised learning, Hsu *et al.* [13] attempt to solve the few-shot problem with unlabeled meta-train data and labeled meta-test data. Different from the above works, we implement meta-learning in the context of fully-unsupervised re-ID, where no labeled data are provided. In addition, this paper adopts the meta-learning to overcome the domain shift in re-ID, considering a completely different problem compared to existing meta-learning methods.

3. Methodology

Preliminary. In the unsupervised person re-ID, we are provided with an unlabeled dataset $U = \{x_1, x_2, \dots, x_N\}$ with N images captured by N_{cam} cameras. Generally, the distributions of images from different cameras will vary greatly. The goal is to learn a model F parameterized by θ with U , which can extract discriminative person feature $f \in \mathbb{R}^{d \times 1}$ for retrieval.

3.1. The Overall Framework

Our overall framework is presented in Fig. 2, which can be summarized into iterations of two stages: *Mining*

Pseudo-Label and Meta-Optimization. In stage one, we first extract features for all training samples and then use DBSCAN [4] to assign pseudo-label for them. In stage two, we aim to use the generated pseudo-label to train the model. Specifically, we maintain a memory W , which saves the features of all samples and is dynamically updated during training. The memory enables us to effectively train the model with changing pseudo-label obtained from the first stage. During training, we split the dataset U into meta-train set M_{tr} and meta-test set M_{te} based on camera labels. The model is trained with our proposed camera-aware meta-learning (MetaCam) strategy, which encourages the model to learn camera-invariant features. In addition, we propose a dynamic and symmetric cross-entropy loss (DSCE) for resisting noisy labels. These two stages are repeatedly iterated till the model converges.

3.2. Mining Pseudo-Label

To enable training on the unlabeled dataset U , this paper adopts a popular way for generating pseudo-label, *i.e.*, clustering-based strategy [30, 34]. Specifically, we first extract pooling-5 features for U by a re-ID model, which is initially pre-trained with ImageNet and will be updated in the training process. Given the extracted features, we compute their pair-wise Euclidean distances and calculate their Jaccard distances with k -reciprocal nearest neighbours [46]. The obtained Jaccard distances are used to generate pseudo-label for U with DBSCAN [4]. Since DBSCAN is a density-based clustering algorithm, it only assigns pseudo-label to high-confident samples (inliers) and remains low-confident samples as outliers. To fully utilize training samples in U , we assign outliers with pseudo-label based on their corresponding nearest neighbours. Based on the above process, we produce pseudo-label for the unlabeled samples, which can be used for model optimization. However, due to the poor model initialization and camera variations, intra-identity samples might be assigned with different pseudo-label and inter-identity samples might be assigned with the same pseudo-label. Training with such noisy labels undoubtedly will hamper the model optimization and thus reduce the model performance. To address this problem, we propose DSCE loss and MetaCam for robust learning and overcoming the camera variations.

3.3. Training with DSCE Loss

Clustering-based unsupervised re-ID largely depends on the iteration between clustering and model optimizing stage [30, 42]. There are two challenges in this process. (1) The number of centroids may change after each iteration, hindering the utilization of traditional cross-entropy loss that requires a fixed number of identities. (2) Clustering algorithm may bring a large amount of noisy samples in both inliers and outliers, which hurts model optimization.

For the first challenge, inspired by [17], we propose a dynamic cross-entropy loss, which can be effectively utilized against the changing of centers. Specifically, we maintain a memory W that saves features for all training samples. During training, we construct online centroids from the memory by directly averaging over memory features that assigned with the same pseudo-label. The dynamic cross-entropy loss is formulated as,

$$L_{dce}(f_i; \cdot) = -\hat{y}_i^T \log \text{Softmax}(C^T f_i / \cdot), \quad (1)$$

where $C \in \mathbb{R}^{N_c \times d}$ represents the feature centers of each pseudo identity, N_c is the number of clusters, d is the feature dimension, and $\text{Softmax}(\cdot)$ is the element-wise softmax function. f_i is the feature of i -th training sample extracted by the current model. $\hat{y}_i \in \mathbb{R}^{N_c \times 1}$ is the one-hot vector indicating the pseudo identity of i -th sample.

For the second challenge, we aim to design a robust loss function for resisting noisy labels. Ghosh *et al.* [9] propose a theory to check whether a loss function L is robust to noisy samples, which can be formulated as:

$$\sum_{k=1}^{N_c} L(f, k) = Z, \quad (2)$$

where N_c is the number of categories and Z is a constant. The above formula indicates that for any sample f and loss function L , the sum of losses about classifying f to all categories (*i.e.*, 1 to N_c) should be a constant if L is noise-tolerant. By utilizing this theory and drawing the inspiration from [36], we introduce a dynamic and symmetric cross-entropy loss (DSCE) as:

$$L_{dsce}(f_i; \cdot) = -\text{Softmax}(C^T f_i / \cdot)^T \log \text{Softmax}(\hat{y}_i). \quad (3)$$

Different from [36], we adopt the softmax normalization to avoid the computational problem brought by $\log 0$ in one-hot vector \hat{y}_i . The proposed L_{dsce} utilizes a memory bank to adapt to the changing clusters in unsupervised re-ID and it also satisfies Eq. 2 (see supplementary material for details). Considering the good convergence of L_{dce} , the combined loss for optimization is:

$$L_c(f_i; \cdot) = L_{dce} + L_{dsce}. \quad (4)$$

After each backpropagation step, we update the feature memory W with the following rule:

$$W[i] = W[i] + (\alpha - \beta) f_i, \quad (5)$$

where $\alpha \in [0, 1]$ is the updating rate.

3.4 Camera-Aware Meta-Learning

The previous training scheme offers a basic solution to unsupervised re-ID, but it ignores the impact of camera shift, which is crucial for optimizing a robust re-ID model. The appearance of pedestrians under different cameras may suffer from the changes of viewpoint, illumination, and other environmental factors, leading to a large gap between intra-class features. Without considering this phenomenon, the trained model might be sensitive to the camera variations, which may decrease the clustering results and thus hampers the model optimization. To address this problem, we propose a camera-aware meta-learning strategy (Meta-Cam) to help the model learn a camera-invariant representation, which includes the following four steps: *Meta-Sets Preparation*, *Meta-Train*, *Meta-Test*, and *Meta-Update*.

Meta-Sets Preparation. In the proposed MetaCam, we aim to align the camera shift by simulating the cross-camera constraint during training. Given the training samples collected from N_{cam} cameras, we split the training set into the meta-train set and meta-test set based on the camera labels. Specifically, we randomly select samples of N_{mtr} cameras as the meta-train set M_{tr} and regard the samples of the rest $N_{\text{cam}} - N_{\text{mtr}}$ cameras as the meta-test set M_{te} . Next, we will introduce how to utilize the generated meta-sets to learn a camera-invariant model.

Meta-Train. We calculate the meta-train loss on the mini-batch examples m_{tr} sampled from M_{tr} with the proposed loss L_c in Eq. 5, formulated as:

$$L_{\text{mtr}}(F(m_{\text{tr}}); \theta) = \frac{1}{N_b} \sum_{i=1}^{N_b} L_c(f_i; \theta), \quad (6)$$

where N_b is the batch size. By updating model parameters with SGD optimizer, we obtain a temporary model parameterized by θ' for further optimization in the *Meta-Test* step. The temporary model is obtained by:

$$\theta' = \theta - \eta \frac{\partial L_{\text{mtr}}}{\partial \theta}, \quad (7)$$

where η is the learning rate.

Meta-Test. In the meta-test step, we aim at validating the accuracy of the temporary model θ' on meta-test samples. To achieve this goal, we sample a mini-batch with N_b images from M_{te} and compute the meta-test loss, formulated as:

$$L_{\text{mte}}(F(m_{\text{te}}); \theta') = \frac{1}{N_b} \sum_{i=1}^{N_b} L_c(f_i; \theta'). \quad (8)$$

Meta-Update. In this step, we update the model with the combination of meta-train loss and meta-test loss, which can be written as:

$$L_{\text{meta}}(F(m_{\text{tr}}), F(m_{\text{te}}); \theta) = L_{\text{mtr}} + L_{\text{mte}}. \quad (9)$$

Algorithm 1 The training procedure of proposed method.

Inputs: Unlabeled data U captured by N_{cam} cameras, batch size N_b , re-ID model F parameterized by θ , feature memory W , number of meta-train cameras N_{mtr} , training epoch epoch , learning rate η , updating rate α .

Outputs: Optimized model F parameterized with θ^* .

```

1: Initialize  $W$  with 0;
2: for  $i$  in epoch do
3:   // Stage 1: Mining Pseudo-Label.
4:   Generate pseudo-label for  $U$  with DBSCAN;
5:   // Stage 2: Meta-Optimization.
6:   // Step 1: Meta-sets Preparation.
7:   Select samples from  $N_{\text{mtr}}$  random cameras as  $M_{\text{tr}}$ 
   and regard samples of the rest cameras as  $M_{\text{te}}$ .
8:   repeat
9:     Sample mini-batch with  $N_b$  meta-train samples
        $m_{\text{tr}}$  and  $N_b$  meta-test samples  $m_{\text{te}}$ .
10:    //Step 2: Meta-Train.
11:    Compute meta-train loss on  $m_{\text{tr}}$  with Eq. 6;
12:    Obtain temporary  $\theta'$  with Eq. 7;
13:    //Step 3: Meta-Test.
14:    Compute meta-test loss on  $m_{\text{te}}$  with Eq. 8;
15:    // Step 4: Meta-Update.
16:    Compute combined loss with Eq. 9;
17:    Update  $\theta$  with gradient computed by Eq. 10;
18:    Update  $W$  with  $m_{\text{tr}}$  and  $m_{\text{te}}$  based on Eq. 5;
19:  until  $M_{\text{tr}}$  and  $M_{\text{te}}$  are enumerated;
20: end for
21:
22: Return  $F$  parameterized with  $\theta^*$ .
```

It should be noted that although the meta-test loss is computed based on temporary model θ' , the derivative *w.r.t.* θ can also be obtained with the chain rule. Specifically, the derivative of L_{meta} *w.r.t.* θ can be formulated as:

$$\frac{\partial L_{\text{meta}}}{\partial \theta} = \frac{\partial L_{\text{mtr}}}{\partial \theta} + \frac{\partial L_{\text{mte}}}{\partial \theta}. \quad (10)$$

To sum up, the overall training process of our method is listed in Alg. 1.

Remark. From Eq. 10, we can observe that the proposed MetaCam encourages the model to be optimized to the direction that can perform well not only on samples from meta-train cameras but also on samples from meta-test cameras. The meta-test loss can be considered as a regularization term, which can lead the model to produce discriminative representations with high-order gradients.

4. Experiments

Datasets and Evaluation Protocol. We evaluate our method on the three large-scale re-ID benchmarks, *i.e.*,

Table 1. Comparison with state-of-the-arts (fully unsupervised). Our method outperforms current unsupervised re-ID algorithms. “**”: Reproduced by [3], “†”: Reproduced based on the authors’ code.

Methods	Venue	DukeMTMC-reID			Market-1501			MSMT-17		
		mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP	rank-1	rank-5
OIM [38]	CVPR’17	11.3	24.5	38.8	14.0	38.0	58.0	-	-	-
BUC [19]	AAAI’19	27.5	47.4	62.6	38.3	66.2	79.6	3.4*	11.5*	18.6*
SSL [20]	CVPR’20	28.6	52.5	63.5	37.8	71.1	83.8	-	-	-
MMCL [34]	CVPR’20	40.2	65.2	75.9	45.5	80.3	89.4	-	-	-
HCT [42]	CVPR’20	50.7	69.6	83.4	56.4	80.0	91.6	-	-	-
ECN [†] [49]	CVPR’19	24.5	49.0	61.7	30.3	63.5	79.0	3.1	10.2	15.5
AE [3]	TOMM’20	39.0	63.2	75.4	54.0	77.5	89.8	8.5	26.6	37.0
WFDR [†] [41]	CVPR’20	42.4	62.0	75.1	50.1	72.1	80.5	8.6	22.3	32.5
Ours	This work	53.8	73.8	84.2	61.7	83.9	92.3	15.5	35.2	48.3

Table 2. Ablation study on the proposed method. “Outliers”: Including outliers into training data. “DSCE”: training with DSCE loss. “MetaCam”: training with MetaCam.

No.	Attributes			DukeMTMC-reID		Market-1501	
	Outliers	DSCE	MetaCam	mAP	rank-1	mAP	rank-1
1	×	×	×	6.8	16.6	6.6	17.5
2		×	×	39.2	59.7	51.2	73.2
3			×	43.4	62.8	53.9	74.8
4		×		51.1	71.2	59.4	82.1
5				53.8	73.8	61.7	83.9

Market-1501 (Market) [44], DukeMTMC-reID (Duke) [26, 45] and MSMT-17 (MSMT) [37]. Market includes 32,668 images from 1,501 persons under six cameras. Duke is composed of 36,411 labeled images of 1,404 identities from eight cameras. MSMT has 126,441 samples from 4,101 pedestrians captured by fifteen cameras. For each dataset, nearly half of the identities are used for training and the remaining identities are used for testing. We adopt mAP and rank-1/5 accuracy for evaluation.

Implementation Details. We adopt the ResNet-50 [12] as the backbone. The “exemplar-invariance” constraint in ECN [49] is used to initialize our model and memory for 5 epochs. In our method, the number of cameras in the meta-train set N_{mtr} are set to 3, 4 and 7 for Market, Duke and MSMT, respectively. During training, we set the learning rate $= 3.5 \times 10^{-4}$, batch size $N_b = 64$, temperature factor $= 0.05$, updating rate $= 0.2$. Images are resized to 256×128 . We use random crop, random flip and random erasing [47] for data augmentation. The model is updated by the Adam optimizer. We train the model with 40 epochs in total, *i.e.*, max_epoch = 40. During testing, we extract 2048-dim pooling-5 features for retrieval.

4.1. Comparison with State-of-the-Art

We evaluate our method on Market, Duke and MSMT and compare it with state-of-the-art methods: including OIM [38], BUC [19], SSL [20], MMCL [34], HCT [42], ECN [49], AE [3] and WFDR [41]. To fairly compare our MetaCam with WFDR [41] that aligns the camera feature shift with 2-Wasserstein distance, we implement WFDR in our framework by replacing MetaCam with WFDR. We

also reproduce ECN [49], which is an unsupervised domain adaptation method and considers the camera shift, based on the provided source code. From Tab. 1, we make the following two conclusions. (1) Our method achieves the best results on three large-scale datasets. Specifically, we achieve **mAP=53.8%** and **rank-1 accuracy=73.8%** for Duke, **mAP=61.7%** and **rank-1 accuracy=83.9%** for Market, and **mAP=15.5%** and **rank-1 accuracy=35.2%** for MSMT. Compared to the currently best published method HCT [42], our method surpasses it by 3.1% on Duke and 5.3% on Market in mAP. This demonstrates that our method produces the new state of the art result for unsupervised person re-ID. (2) Compared to methods (ECN [49], AE [3], and WFDR [41]) that consider the camera variations during model training, our method produces significantly higher results. Specifically, when using the same framework, our method (w/ MetaCam) clearly outperforms WFDR [41] in all datasets. This demonstrates the effectiveness of the proposed MetaCam in addressing the camera shift for unsupervised re-ID.

4.2. Ablation Study

We conduct experiments to investigate three important components of our methods, *i.e.* 1) using outliers that assign labels for low-confident samples generated by clustering, 2) DSCE loss that is designed to prevent the model from overfitting on noisy samples, and 3) MetaCam that is proposed to overcome camera variations. Ablative experiments on these three components are reported in Tab. 2.

The effectiveness of using outliers. Without using these three components, the model achieves poor results on both datasets. The main reason is that DBSCAN will regard most of the samples as outliers with features extracted by the poor initial model, leading to the limited training samples during optimization and thereby produces undesired results. When using outliers during training, the results are significantly improved. This demonstrates the importance of assigning pseudo-label for outliers and using them during model training. In the following experiments, we use outliers in training by default.

The effectiveness of DSCE loss function. When adding

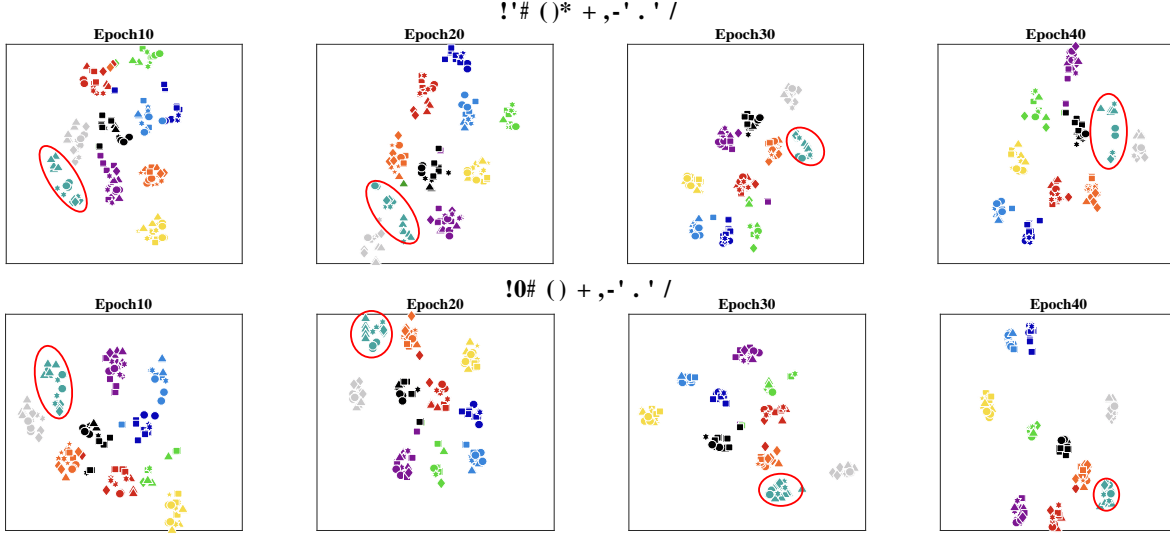


Figure 3. t -SNE plot of 10 persons under different settings (model trained w/o MetaCam and model trained w/ MetaCam). We use different colors to denote identities and different shapes to indicate camera IDs. The algorithm with MetaCam generates intra-class features that are close to each other, indicating that our MetaCam can guide the model to learn camera-invariant features.

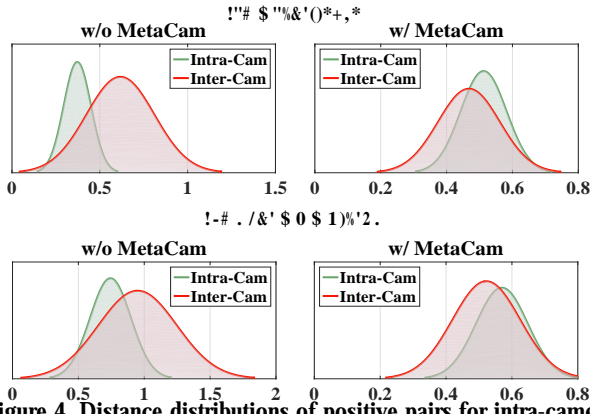


Figure 4. Distance distributions of positive pairs for intra-camera (green curve) and inter-camera (red curve).

DSCE loss into the model, we can achieve consistent improvement, no matter whether to use MetaCam or not. This verifies the advantage of DSCE loss when training the model with noisy labels generated by the clustering step.

The effectiveness of MetaCam. From the comparison of No. 2 vs No. 4 and No. 3 vs No. 5, we obtain two observations. First, MetaCam can significantly improve the results, demonstrating the necessity of overcoming the camera variation in unsupervised re-ID and the effectiveness of the proposed MetaCam. Second, the proposed MetaCam and DSCE loss are complementary to each other. When combining them, the model can gain more improvement in performance.

4.3. Visualization for MetaCam

To better understand the effect of our MetaCam in overcoming the camera variations, we conduct two visualization

experiments: (1) t -SNE [21] plot of feature embeddings with the evolution of training; (2) distance distribution of intra-camera and inter-camera samples for the same ID.

t -SNE plot of feature embeddings. We randomly select 10 persons from Market and visualize their features with t -SNE [21] in different training epochs. In Fig. 3, we show the results of the model trained with or without MetaCam, respectively. For fair comparison, both models use the DSCE loss during training. We use different colors to denote identities and different shapes to represent camera IDs. We can find two phenomena. (1) Features of the same identity are progressively gathered with the model training for both settings. This demonstrates that our method is able to learn discriminative person representations. (2) The model trained with MetaCam can produce more compact feature clusters (*e.g.*, dark green points highlighted by the red circle). The intra-class features under different cameras are well gathered with the help of MetaCam. This verifies the advantage of our MetaCam in learning camera-invariant features. In addition, with the camera-invariant features, we can generate more accurate pseudo-label in the clustering step, which can further facilitate the optimization.

Distance distribution. To more precisely investigate the influence of MetaCam, we conduct experiments to visualize the distance distribution of positive pairs for intra-camera and inter-camera. Specifically, we randomly select 50,000 distances for both intra-camera pairs and inter-camera pairs and draw the histogram for each setting. Results compared between the model trained with or without MetaCam are illustrated in Fig. 4. We can make two observations. (1) The model trained without MetaCam leads to a large gap between intra-camera distribution and inter-camera distri-

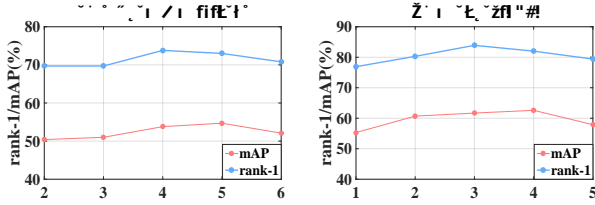


Figure 5. Sensitivity analysis of N_{mtr} .

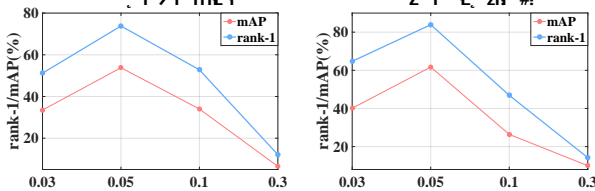


Figure 6. Sensitivity analysis of τ .

bution. Specifically, the distances between positive pairs of inter-camera are commonly larger than that of intra-camera, indicating that the model trained without MetaCam is sensitive to camera variations. (2) When training with MetaCam, the distribution gap between positive pairs for intra-camera and inter-camera is significantly reduced. Concretely, the distances between positive pairs of inter-camera are commonly similar to that of intra-camera. This suggests that our MetaCam is able to align the camera shift and can lead the model to learn camera-invariant features, which is an important factor in person re-ID.

4.4. Sensitivity Analysis

We further analyze the sensitivity to two hyper-parameters of our method, *i.e.*, the number of cameras in meta-train N_{mtr} for MetaCam and temperature factor for the memory-based loss. In our experiments, we change the value of one hyper-parameter and the others remain fixed.

Sensitivity to N_{mtr} . In Fig. 5, we vary N_{mtr} from 1 to 5 for Market and 2 to 6 for Duke to investigate the effect of involving samples of different cameras into meta-train. We find that the best accuracy is achieved when N_{mtr} is equal to half of the total number of cameras for each dataset. This indicates that it is better to keep a balance camera variations in the meta-train and meta-test.

Sensitivity to τ . In Fig. 6, we investigate the effect of temperature factor τ . A smaller τ leads to lower entropy and may help to achieve better results in re-ID. However, the τ with too small value, such as 0.03, will cause the collapse of training. In our experiments, we set $\tau = 0.05$, which achieves well performance across all datasets.

4.5. Results for Domain Adaptation

We also apply our method to the setting of unsupervised domain adaptation (UDA). In UDA, we are additionally given a labeled source domain, which can provide extra su-

Table 3. Results on domain adaptation. M: Market-1501, D: DukeMTMC-reID. MMT-500: MMT [7] with $k = 500$ for k -means clustering. All methods use ResNet-50 as the backbone.

Methods	Venue	D		M	
		mAP	rank-1	mAP	rank-1
SPGAN [2]	CVPR'18	22.8	51.5	22.3	44.1
HHL [48]	ECCV'18	31.4	62.2	27.2	46.9
ECN [49]	CVPR'19	43.0	75.1	40.4	63.3
SSG [6]	ICCV'19	58.3	80.0	53.4	73.0
UCDA-CCE [24]	ICCV'19	34.5	64.3	36.7	55.4
MMCL [34]	CVPR'20	60.4	84.4	51.4	72.4
DG-Net++ [52]	ECCV'20	61.7	82.1	63.8	78.9
GDS [14]	ECCV'20	61.2	81.1	55.1	73.1
ECN+ [50]	TPAMI'20	63.8	84.1	54.4	74.0
MMT-500 [7]	ICLR'20	71.2	87.7	63.1	76.8
MMT-500+Ours	This Work	76.5	90.1	65.0	79.5

pervision for model training. Since our method is designed to learn the model with unlabeled data, we initialize the re-ID backbone with the model trained on MMT [7] and fine-tune the model with our method on the unlabeled data. We evaluate our method on the settings of transferring between Duke and Market. Comparisons with state-of-the-art methods are reported in Tab. 3. All the compared methods use ResNet-50 as the backbone. We can observe that MMT [7] achieves the best results. After adding our method, the performance is further improved. Specifically, our method increases the mAP from 71.2% to 76.5% when testing on Market and from 63.1% to 65.0% when testing on Duke. This indicates that our method is also suitable for UDA and can be readily applied to further improve the performance of other UDA methods.

5. Conclusion

In this paper, we propose a novel framework for unsupervised re-ID, which is designed based on a Dynamic and Symmetric Cross-Entropy loss (DSCE) and a camera-aware meta-learning algorithm (MetaCam). Our DSCE is able to handle the changing clusters and can resist noisy samples during model optimization. The proposed MetaCam can effectively reduce the camera shift by simulating the cross-camera searching process during training. Extensive experiments show the effectiveness of our method, which can achieve state-of-the-art results on three datasets for both fully-unsupervised re-ID and domain adaptive re-ID.

Acknowledgements This work is supported by the National Nature Science Foundation of China (No. 61876159, 61806172, 61662024, 62076116 and U1705286); the China Postdoctoral Science Foundation Grant (No. 2019M652257); the Fundamental Research Funds for the Central Universities (Xiamen University, No. 20720200030); the EU H2020 AI4Media No. 951911 project; and the Italy-China collaboration project TALENT:2018YFE0118400.

References

- [1] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *ICLR*, 2021. **3**
- [2] 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**
- [3] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. *ACM TOMM*, 16(1), 2020. **6**
- [4] 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. **3, 4**
- [5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. **3**
- [6] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, 2019. **2, 8**
- [7] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. **1, 2, 8**
- [8] 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**
- [9] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017. **3, 4**
- [10] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *CVPR*, 2020. **3**
- [11] 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**
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. **6**
- [13] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *ICLR*, 2019. **3**
- [14] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Global distance-distributions separation for unsupervised person re-identification. In *ECCV*, 2020. **8**
- [15] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, 2018. **3**
- [16] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. **3**
- [17] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. **4**
- [18] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *ICCV*, 2019. **2**
- [19] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019. **1, 2, 6**
- [20] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *CVPR*, 2020. **1, 2, 6**
- [21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov), 2008. **7**
- [22] Tsendsuren Munkhdalai and Hong Yu. Meta networks. *ICML*, 2017. **3**
- [23] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2018. **3**
- [24] 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. **2, 8**
- [25] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *NeurIPS*, 2019. **3**
- [26] 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*, 2016. **6**
- [27] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016. **3**
- [28] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. **3**
- [29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. **3**
- [30] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 102, 2020. **4**
- [31] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. **1, 3**
- [32] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. **3**
- [33] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *NeurIPS*, 2017. **3**
- [34] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, 2020. **2, 4, 6, 8**
- [35] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018. **1**

- [36] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019. 3, 4
- [37] 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, 6
- [38] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017. 6
- [39] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 3
- [40] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *AAAI*, 2020. 3
- [41] Hong-Xing Yu and Wei-Shi Zheng. Weakly supervised discriminative feature learning with state information for person identification. In *CVPR*, 2020. 2, 3, 6
- [42] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *CVPR*, 2020. 1, 2, 4, 6
- [43] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018. 3
- [44] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *CVPR*, 2015. 6
- [45] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 6
- [46] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 4
- [47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 6
- [48] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, 2018. 2, 8
- [49] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 1, 2, 6, 8
- [50] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *TPAMI*, 2020. 8
- [51] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *TIP*, 2019. 3
- [52] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, 2020. 2, 8