Rethinking the Distribution Gap of Person Re-identification with Camera-based Batch Normalization

Zijie Zhuang[1], Longhui Wei[2,4], Lingxi Xie[2], Tianyu Zhang[2], Hengheng Zhang[3], Haozhe Wu[1], Haizhou Ai[1], and Qi Tian[2]

[1]Tsinghua University [2]Huawei Inc.
[3]Hefei University of Technology [4]University of Science and Technology of China
{jayzhuang42,weilh2568,198808xc,tianyu1949,imhmhm}@gmail.com
wuhz1997@163.com, ahz@tsinghua.edu.cn, tian.qi1@huawei.com

Abstract. The fundamental difficulty in person re-identification (ReID) lies in learning the correspondence among individual cameras. It strongly demands costly inter-camera annotations, yet the trained models are not guaranteed to transfer well to previously unseen cameras. These problems significantly limit the application of ReID. This paper rethinks the working mechanism of conventional ReID approaches and puts forward a new solution. With an effective operator named Camera-based Batch Normalization (CBN), we force the image data of all cameras to fall onto the same subspace, so that the distribution gap between any camera pair is largely shrunk. This alignment brings two benefits. First, the trained model enjoys better abilities to generalize across scenarios with unseen cameras as well as transfer across multiple training sets. Second, we can rely on intra-camera annotations, which have been undervalued before due to the lack of cross-camera information, to achieve competitive ReID performance. Experiments on a wide range of ReID tasks demonstrate the effectiveness of our approach. The code is available at https://github.com/automan000/Camera-based-Person-ReID.

Keywords: Person Re-identification, Distribution Gap, Camera-based Batch Normalization

1 Introduction

Person re-identification (ReID) aims at matching identities across disjoint cameras. Generally, it is achieved by mapping images from the same and different cameras into a feature space, where features of the same identity are closer than those of different identities. To learn the relations between identities from all cameras, there are two different objectives: learning the relations between identities in the same camera and learning identity relations across cameras.

However, there is an inconsistency between these two objectives. As shown in Fig. 1(a), due to the large appearance variation caused by illumination conditions, camera views, *etc.*, images from different cameras are subject to distinct

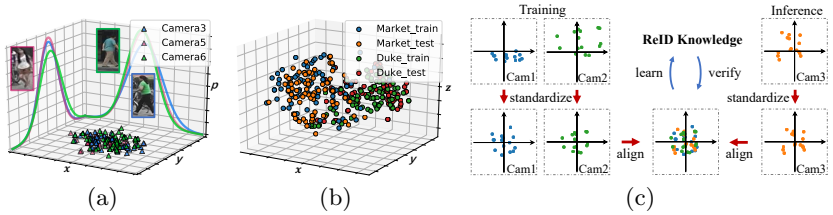


Fig. 1. (a) We visualize the distributions of several cameras in Market-1501. Each curve corresponds to an approximated marginal density function. Curves of different cameras demonstrate the differences between the corresponding distributions. (b) The Barnes-Hut t-SNE [40] visualization of the distribution inconsistency among datasets. (c) Illustration of the proposed camera-based formulation. Note that **Cam1**, **Cam2**, and **Cam3** could come from any ReID datasets. This figure is best viewed in color.

distributions. Handling the distribution gap between cameras is crucial for inter-camera identity matching, yet learning within a single camera is much easier. As a consequence, the conventional ReID approaches mainly focus on associating different cameras, which demands costly inter-camera annotations. Besides, after learning on a training set, part of the learned knowledge is strongly correlated to the connections among these particular cameras, making the model generalize poorly on scenarios consisting of unseen cameras. As shown in Fig. 1(b), the ReID model learned on one dataset often has a limited ability of describing images from other datasets, *i.e.*, its generalization ability across datasets is limited. For simplicity, we denote this formulation neglecting within-dataset inconsistencies as the **dataset-based formulation**. We emphasize that lacking the ability to bridge the distribution gap between all cameras from all datasets leads to two problems: the unsatisfying generalization ability and the excessive dependence on inter-camera annotations. To tackle these problems simultaneously, we propose to align the distribution of all cameras explicitly. As shown in Fig. 1(c), we eliminate the distribution inconsistency between all cameras, so the ReID knowledge can always be learned, accumulated, and verified in the same input distribution, which facilitates the generalization ability across different ReID scenarios. Moreover, with the aligned distributions among all cameras, intra- and inter-camera annotations can be regarded as the same, *i.e.*, labeling the image relations under the same input distribution. This allows us to approximate the effect of inter-camera annotations with only intra-camera annotations. It may relieve the exhaustive human labor for the costly inter-camera annotations.

We denote our solution that disassembles ReID datasets and aligns each camera independently as the **camera-based formulation**. We implement it via an improved version of Batch Normalization (BN) [9] named Camera-based Batch Normalization (CBN). In training, CBN disassembles each mini-batch and standardizes the corresponding input according to its camera labels. In testing, CBN utilizes few samples to approximate the BN statistics of every testing camera and standardizes the input to the training distribution. In practice, multiple ReID

tasks benefit from our work, such as *fully-supervised learning* [1,52,37,54,55,59], *direct transfer* [22,8], *domain adaptation* [42,3,58,4,34,53], and *incremental learning* [29,16,12]. Extensive experiments indicate that our method improves the performance of these tasks simultaneously, such as 0.9%, 5.7%, and 14.2% averaged Rank-1 accuracy improvements on *fully-supervised learning*, *domain adaptation*, and *direct transfer*, respectively, and 9.7% less forgetting on Rank-1 accuracy for *incremental learning*. Last but not least, even without inter-camera annotations, a *weakly-supervised* pipeline [61] with our formulation can achieve competitive performance on multiple ReID datasets, which demonstrates that the value of intra-camera annotations may have been undervalued in the previous literature. To conclude, our contribution is three-fold:

- In this paper, we emphasize the importance of aligning the distribution of all cameras and propose a camera-based formulation. It can learn discriminative knowledge for ReID tasks while excluding training-set-specific information.
- We implement our formulation with Camera-based Batch Normalization. It facilitates the generalization and transfer ability of ReID models across different scenarios and makes better use of intra-camera annotations. It provides a new solution for ReID tasks without costly inter-camera annotations.
- Experiments on *fully-supervised*, *weakly-supervised*, *direct transfer*, *domain adaptation*, and *incremental learning* tasks validate our method, which confirms the universality and effectiveness of our camera-based formulation.

2 Related Work

Our formulation aligns the distribution per camera. In training, it eliminates the distribution gap between all cameras. ReID models can treat intra- and inter-camera annotations equally and make better use of them, which benefits both *fully-supervised* and *weakly-supervised* ReID tasks. It also guarantees that the distribution of each testing camera is aligned to the same training distribution. Thus, the knowledge can better generalize and transfer across datasets. It helps *direct transfer*, *domain adaptation*, and *incremental learning*. In this section, we briefly categorize and summarize previous works on the above ReID topics.

Supervision. The supervision in ReID tasks is usually in the form of identity annotations. Although there are many outstanding unsupervised methods [46,45,48,47] that do not need annotations, it is usually hard for them to achieve competitive performance as the supervised ReID methods. For better performance, lots of previous methods [1,52,37,54,55,59,11,43] utilized *fully-supervised learning*, in which identity labels are annotated manually across all training cameras. Many of them designed spatial alignment [50,38,35], visual attention [13,20], and semantic segmentation [11,39,32] for extracting accurate and fine-grained features. GAN-based methods [21,10,24] were also utilized for data augmentation. However, although these methods achieved remarkable performance on ReID tasks, they required costly inter-camera annotations. To reduce the cost of human labor, ReID researchers began to investigate *weakly-supervised learning*. SCT [49] presumes that each identity appears in only one

camera. In ICS [61], an intra-camera supervision task is studied in which an identity could have different labels under different cameras. In [18,19], pseudo labels are used to supervised the ReID model.

Generalization. The generalization ability in ReID tasks denotes how well a trained model functions on unseen datasets, which is usually examined by *direct transfer* tasks. Researchers found that many fully-supervised ReID models perform poorly on unseen datasets [33,42,3]. To improve the generalization ability, various strategies were adopted as additional constraints to avoid over-fitting, such as label smoothing [22] and sophisticated part alignment approaches [8].

Transfer. The transfer ability in ReID tasks corresponds to the capability of ReID models transferring and preserving the discriminative knowledge across multiple training sets. There are two related tasks. *Domain adaptation* transfers knowledge from labeled source domains to unlabeled target domains. One solution [42,3,58] bridged the domain gap by transferring source images to the target image style. Other solutions [6,41,4,17,34] utilized the knowledge learned from the source domain to mine the identity relations in target domains. *Incremental learning* [29,16,12] also values the transfer ability. Its goal is to preserve the previous knowledge and accumulate the common knowledge for all seen datasets. A recent ReID work that relates to incremental learning is MASDF [44], which distilled and incorporated the knowledge from multiple datasets.

3 Methodology

3.1 Conventional ReID: Learning Camera-related Knowledge

ReID is a task of retrieving identities according to their appearance. Given a training set consisting of disjoint cameras, learning a ReID model on it requires two types of annotations: inter-camera annotations and intra-camera annotations. The conventional ReID formulation regards a ReID dataset as a whole and learns the relations between identities as well as the connections between training cameras. Given an image $\mathbf{I}_i^{\mathcal{D}_j}$ from any training set $\mathcal{D}_j$, the training goal of this formulation is:

$$\arg\min \mathbb{E}\left[\mathbf{y}_i^{\mathcal{D}_j} - \mathbf{g}^{\mathcal{D}_j}\left(\mathbf{f}^{\mathcal{D}_j}\left(\mathbf{I}_i^{\mathcal{D}_j}\right)\right)\right], \left(\mathbf{I}_i^{\mathcal{D}_j}, \mathbf{y}_i^{\mathcal{D}_j}\right) \in \mathcal{D}_j, \tag{1}$$

where $\mathbf{f}^{\mathcal{D}_j}(\cdot)$ and $\mathbf{g}^{\mathcal{D}_j}(\cdot)$ are the corresponding feature extractor and classifier for $\mathcal{D}_j$, respectively. $\mathbf{y}_i^{\mathcal{D}_j}$ denotes the identity label of the image $\mathbf{I}_i^{\mathcal{D}_j}$.

In our opinion, this formulation has three drawbacks. First, images from different cameras, even of the same identity, are subject to distinct distributions. To associate images across cameras, conventional approaches strongly demand the costly inter-camera annotations. Meanwhile, the intra-camera annotations are less exploited since they provide little information across cameras. Second, such learned knowledge not only discriminates the identities in the training set but also encodes the connections between training cameras. These connections are associated with the particular training cameras and hard to generalize to

other cameras, since the learned knowledge may not apply to the distribution of previously unseen cameras. For example, when transferring a ReID model trained on Market-1501 to DukeMTMC-reID, it produces a poor Rank-1 accuracy of 37.0% without fine-tuning. Third, the learned knowledge is hard to preserve when being fine-tuned. For instance, after fine-tuning the aforementioned model on DukeMTMC-reID, the Rank-1 accuracy drops 14.2% on Market-1501, because it turns to fit the relations between the cameras in DukeMTMC-reID. We analyze these three problems and find that the particular relations between training cameras are the primary cause of them. Thus, we believe that the conventional method of handling these camera-related relations may need a re-design.

3.2 Our Insight: Towards Camera-independent ReID

We rethink the relations between cameras. More specifically, we believe that the exclusive knowledge for bridging the distribution gap between the particular training cameras should be suppressed during training. Such knowledge is associated to the cameras in the training set and sacrifices the discriminative and generalization ability on unseen scenarios.

To this end, we propose to align the distribution of all cameras explicitly, so that the distribution gap between all cameras is eliminated, and much less camera-specific knowledge will be learned during training. We denote this formulation as the **camera-based formulation**. To align the distribution of each camera, we estimate the raw distribution of each camera and standardize images from each camera with the corresponding distribution statistics. We use $\boldsymbol{\eta}(\cdot)$ to denote the estimated statistics related to the distribution of a camera. Then, given a related image $\mathbf{I}_i^{(c)}$, aligning the camera-wise distribution will transform this image as:

$$\tilde{\mathbf{I}}_i^{(c)} = \mathbf{DA}\left(\mathbf{I}_i^{(c)}; \boldsymbol{\eta}(c)\right), \tag{2}$$

where $\mathbf{DA}(\cdot)$ represents a distribution alignment mechanism, $\tilde{\mathbf{I}}_i^{(c)}$ denotes the aligned $\mathbf{I}_i^{(c)}$ and $\boldsymbol{\eta}(c)$ is the estimated alignment parameters for camera c. For any training set $\mathcal{D}_j$, we can now learn the ReID knowledge from this aligned distribution by replacing $\mathbf{I}_i^{\mathcal{D}_j}$ in Eq. 1 with $\tilde{\mathbf{I}}_i^{(c)}$.

With the distributions of all cameras aligned by $\mathbf{DA}(\cdot)$, images from all these cameras can be regarded as distributing on a "standardized camera". By learning on this "standardized camera", we eliminate the distribution gap between cameras, so the raw learning objectives within the same and across different cameras can be treated equally, making the training procedure more efficient and effective. Besides, without the disturbance caused by the training-camera-related connections, the learned knowledge can generalize better across various ReID scenarios. Last but not least, now that the additional knowledge for associating diverse distributions is much less required, our formulation can make better use of the intra-camera annotations. It may relieve human labor for the costly inter-camera annotations, and provides a solution for ReID in a large-scale camera network with fewer demands of inter-camera annotations.

3.3 Camera-based Batch Normalization

In practice, a possible solution for aligning camera-related distributions is to conduct batch normalization in a camera-wise manner. We propose the Camera-based Batch Normalization (CBN) for aligning the distribution of all training and testing cameras. It is modified from the conventional Batch Normalization [9], and estimates camera-related statistics rather than dataset-related statistics.

Batch Normalization Revisited. The Batch Normalization [9] is designed to reduce the internal covariate shifting. In training, it standardizes the data with the mini-batch statistics and records them for approximating the global statistics. During testing, given an input $\mathbf{x}_i$, the output of the BN layer is:

$$\hat{\mathbf{x}}_i = \gamma \frac{\mathbf{x}_i - \hat{\mu}}{\sqrt{\hat{\sigma}^2 + \epsilon}} + \beta, \tag{3}$$

where $\mathbf{x}_i$ is the input and $\hat{\mathbf{x}}_i$ is the corresponding output. $\hat{\mu}$ and $\hat{\sigma}^2$ are the global mean and variance of the training set. γ and β are two parameters learned during training. In ReID tasks, BN has significant limitations. It assumes and requires that all testing images are subject to the same training distribution. However, this assumption is satisfied only when the cameras in the testing set and training set are exactly the same. Otherwise, the standardization fails.

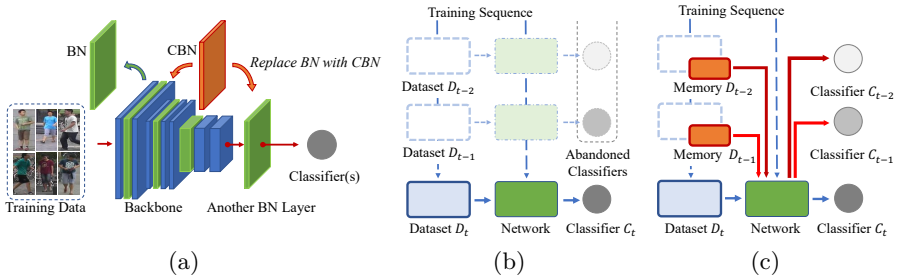
Batch Normalization within Cameras. Our Camera-based Batch Normalization (CBN) aligns all training and testing cameras independently. It guarantees an invariant input distribution for learning, accumulating, and verifying the ReID knowledge. Given images or corresponding intermediate features $\mathbf{x}_m^{(c)}$ from camera c, CBN standardizes them according to the camera-related statistics:

$$\mu_{(c)} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}_m^{(c)}, \quad \sigma_{(c)}^2 = \frac{1}{M}\sum_{m=1}^{M} \left(\mathbf{x}_m^{(c)} - \mu_{(c)}\right)^2, \quad \hat{\mathbf{x}}_m = \gamma \frac{\mathbf{x}_m - \mu_{(c)}}{\sqrt{\sigma_{(c)}^2 + \epsilon}} + \beta, \tag{4}$$

where $\mu_{(c)}$ and $\sigma_{(c)}^2$ denote the mean and variance related to this camera c. During training, we disassemble each mini-batch and calculate the camera-related mean and variance for each involved camera. The camera with only one sampled images is ignored. During testing, before employing the learned ReID model to extract features, the above statistics have to be renewed for every testing camera. In short, we collect several unlabeled images and calculate the camera-related statistics per testing camera. Then, we employ these statistics and the learned weights to generate the final features.

3.4 Applying CBN to Multiple ReID Scenarios

The proposed CBN is generic and nearly cost-free for existing methods on multiple ReID tasks. To demonstrate its superiority, we setup a bare-bones baseline, which only contains a deep neural network, an additional BN layer as the bottleneck, and a fully connected layer as the classifier. As shown in Fig. 2(a), our



(a) (b) (c)

Fig. 2. Demonstrations of our bare-bones baseline network and two *incremental learning* settings involved in this paper. (a) Given an arbitrary backbone with BN layers, we simply replace all BN layers with our CBN layers. (b) **Data-Free**. (c) **Replay**.

camera-based formulation can be implemented by simply replacing all BN layers in a usual convolutional network with CBN layers.

With a modified network mentioned above, our camera-based formulation can be applied to many popular tasks, such as *fully-supervised learning*, *weakly-supervised learning*, *direct transfer*, and *domain adaptation*. Apart from them, we also evaluate a rarely discussed ReID task, *i.e.*, *incremental learning*. It studies the problem of learning knowledge incrementally from a sequence of training sets while preserving and accumulating the previously learned knowledge. As shown in Fig. 2, we propose two settings. (1) **Data-Free**: once we finish the training procedure on a dataset, the training data along with the corresponding classifier are abandoned. When training the model on the subsequent training sets, the old data will never show up again. (2) **Replay**: unlike Data-Free, we construct an exemplar set from each old training set. The exemplar set and the corresponding classifier are preserved and used during the entire training sequence.

3.5 Discussions

Bridging ReID Tasks. We briefly demonstrate our understandings of the relations between ReID tasks and how we bridge these tasks. Different ReID tasks handle different combinations of training and testing sets. Since datasets have distinct cameras, previous methods have to learn exclusive relations between particular training cameras and adapt them to specific testing camera sets. Our formulation aligns the distribution of all cameras for learning and testing ReID knowledge, and suppresses the exclusive training-camera relations. It may reveal the latent connections between ReID tasks. First, by aligning the distribution of seen and unseen cameras, *fully-supervised learning* and *direct transfer* are united since training and testing distributions are always aligned in a camera-wise manner. Second, since there is no need to learn relations between distinct camera-related distributions, intra- and inter-camera annotations can be treated almost equally. Knowledge is better shared among cameras which helps *fully-* and *weakly-supervised learning*. Third, with the aligned training and testing distri-

butions, it is more efficient to learn, accumulate, and preserve knowledge across datasets. It offers an elegant solution to preserve old knowledge (*incremental learning*) and absorb new knowledge (*domain adaptation*) in the same model.

Relationship to Previous Works. There are two types of previous works that closely relate to ours: camera-related methods and BN variants. Camera-related methods such as CamStyle [58] and CAMEL [46] noticed the camera view discrepancy inside the dataset. CamStyle augmented the dataset by transferring the image style in a camera-to-camera manner, but still learned ReID models in the dataset-based formulation. Consequently, transferring across datasets is still difficult. CAMEL [46] is the most similar work with ours, which learned camera-related projections and mapped camera-related distributions into an implicit common distribution. However, these projections are associated with the training cameras, limiting its ability to transfer across datasets. BN variants such as AdaBN also inspire us. AdaBN aligned the distribution of the entire dataset. It neither eliminated the camera-related relations in training nor handled the camera-related distribution gap in testing. Unlike them, CBN is specially designed for our camera-based formulation. It is much more general and precise for ReID tasks. More comparisons will be provided in Secs. 4.2 and 4.3.

4 Experiments

4.1 Experiment Setup

Datasets. We utilize three large scale ReID datasets, including Market-1501 [51], DukeMTMC-reID [53], and MSMT17 [42]. Market-1501 dataset has 1,501 identities in total. 751 identities are used for training and the rest for testing. The training set contains 12,936 images and the testing set contains 15,913 images. DukeMTMC-reID dataset contains 16,522 images of 702 identities for training, and 1,110 identities with 17,661 images are used for testing. MSMT17 dataset is the current largest ReID dataset with 126,441 images of 4,101 identities from 15 cameras. For short, we denote Market-1501 as Market, DukeMTMC-reID as Duke, and MSMT17 as MSMT in the rest of this paper. *It is worth noting that in these datasets, the training and testing subsets contain the same camera combinations. It could be the reason that previous dataset-based methods create remarkable fully-supervised performance but catastrophic direct transfer results.*

Implementation Details. In this paper, all experiments are conducted with PyTorch. The image size is 256×128 and the batch size is 64. In training, we sample 4 images for each identity. The baseline network presented in Sec. 3.4 uses ResNet-50 [7] as the backbone. To train this network, we adopt SGD optimizer with momentum [28] of 0.9 and weight decay of 5×10^{-4}. The initial learning rate is 0.01, and it decays after the $40th$ epoch by a factor of 10. For all experiments, the training stage will end up with 60 epochs. For incremental learning, we include a warm-up stage. In this stage, we freeze the backbone and only fine-tune the classifier(s) to avoid damaging the previously learned knowledge. During testing, our framework will first sample a few unlabeled images from each camera and use them to approximate the camera-related statistics. Then,

Table 1. Results of the baseline method with our formulation and the conventional formulation. The fully-supervised learning results are in *italics*.

Training Set	Testing Set	Market		Duke		MSMT	
	Formulation	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Market	Conventional	*90.2*	*74.0*	37.0	20.7	17.1	5.5
	Ours	***91.3***	***77.3***	**58.7**	**38.2**	**25.3**	**9.5**
Duke	Conventional	53.2	25.1	*81.5*	*66.6*	27.2	9.1
	Ours	**72.7**	**43.0**	***82.5***	***67.3***	**35.4**	**13.0**
MSMT	Conventional	58.1	30.8	57.8	38.4	*71.5*	*42.3*
	Ours	**73.7**	**45.0**	**66.2**	**46.7**	***72.8***	***42.9***

these statistics are fixed and employed to process the corresponding testing images. Following the conventions, mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) curves are utilized for evaluations.

4.2 Performance on Different ReID Tasks

We evaluate our proposed method on five types of ReID tasks, *i.e.*, *fully-supervised learning*, *weakly-supervised learning*, *direct transfer*, *domain adaptation*, and *incremental learning*. The corresponding experiments are organized as follows. First, we demonstrate the importance of aligning the distribution of all cameras from all datasets, and simultaneously conduct *fully-supervised learning* and *direct transfer* on multiple ReID datasets. Second, we demonstrate that it is possible to learn discriminative knowledge with only intra-camera annotations. We utilize the network architecture in Sec. 3.4 to compare the *fully-supervised learning* and *weakly-supervised learning*. To evaluate the generalization ability, *direct transfer* is also conducted for these two settings. Third, we evaluate the transfer ability of our method. This part of experiments includes *domain adaptation*, *i.e.*, transferring the knowledge from the old domain to new domains, and *incremental learning*, *i.e.*, preserving the old knowledge and accumulating the common knowledge for all training sets.

Note that, for simplicity, we denote the results of training and testing the model on the same dataset with fully annotated data as the *fully-supervised learning results*. For similar experiments that only use the intra-camera annotations, we denote their results as the *weakly-supervised learning results*.

Supervisions and Generalization. In this section, we evaluate and analyze the supervisions and the generalization ability in ReID tasks. For all experiments in this section, the testing results on both the training domain and other unseen testing domains are always obtained by the same learned model. We first conduct experiments on *fully-supervised learning* and *direct transfer*. As shown in Tab. 1, our proposed method shows good advantages, *e.g.*, there is an averaged 1.1% improvement in Rank-1 accuracy for the *fully-supervised learning* task. Meanwhile, without bells and whistles, there is an average 13.6% improvement in Rank-1 accuracy for the *direct transfer* task. We recognize that our method

Table 2. Results of the state-of-the-art fully-supervised learning methods. BoT* denotes our results with the official BoT code. In BoT*, Random Erasing is disabled due to its negative effect on direct transfer. Unless otherwise stated, the **baseline** method in the following sections refers to the network described in Sec. 3.4.

Method	Market				Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
CamStyle [58]	88.1	-	-	68.7	75.3	-	-	53.5
MLFN [2]	90.0	-	-	74.3	81.0	-	-	62.8
SCPNet [5]	91.2	97.0	-	75.2	80.3	89.6	-	62.6
HA-CNN [14]	91.2	-	-	75.7	80.5	-	-	63.8
PGFA [25]	91.2	-	-	76.8	82.6	-	-	65.5
MVP [36]	91.4	-	-	80.5	83.4	-	-	70.0
SGGNN [31]	92.3	96.1	97.4	82.8	81.1	88.4	91.2	68.2
SPReID [11]	92.5	97.2	98.1	81.3	84.4	91.9	93.7	71.0
BoT* [22]	93.6	97.6	98.4	82.2	84.3	91.9	94.2	70.1
PCB+RPP [38]	93.8	97.5	98.5	81.6	83.3	90.5	92.5	69.2
OSNet [60]	94.8	-	-	84.9	88.6	-	-	73.5
VA-reID [62]	**96.2**	**98.7**	-	**91.7**	**91.6**	**96.2**	-	**84.5**
Baseline	90.2	96.7	97.9	74.0	81.5	91.4	94.0	66.6
Ours+Baseline	91.3	97.1	98.4	77.3	82.5	91.7	94.1	67.3
Ours+BoT*	94.3	97.9	98.7	83.6	84.8	92.5	95.2	70.1

has to collect a few unlabeled samples from each testing camera for estimating the camera-related statistics. However, this process is fast and nearly cost-free.

Our method can also boost previous methods. Take BoT [22], a recent state-of-the-art method, as an example. We integrate our proposed CBN into BoT and conduct experiments with almost the same settings as in the original paper, including the network architecture, objective functions, and training strategies. The only difference is that we disable Random Erasing [55] due to its constant negative effects on *direct transfer*. The results of the *fully-supervised learning* on Market and Duke are shown in Tab. 2. It should be pointed out that in *fully-supervised learning*, training and testing subsets contain the same cameras. Therefore, there is no significant shift among the BN statistics of the training set and the testing set, which favors the conventional formulation. Even so, our method still improves the performance on both Market and Duke. We believe that both aligning camera-wise distributions and better utilizing all annotations contribute to these improvements. Moreover, we also present results on *direct transfer* in Tab. 4. It is clear that our method improves BoT significantly, *e.g.*, there is a 15.3% Rank-1 improvement when training on Duke but testing on Market. These improvements on both *fully-supervised learning* and *direct transfer* demonstrate the advantages of our camera-based formulation.

Weak Supervisions. As we demonstrated in Sec. 3.1, the conventional ReID formulation strongly demands the inter-camera annotations for associating identities under distinct camera-related distributions. Since our method eliminates the distribution gap between cameras, the intra-camera annotations can be bet-

Table 3. The comparisons of fully- and weakly-supervised learning. Results of training and testing on the same domain are in *italics*. MT [61] is our baseline. Except for the camera-based formulation, our weakly-supervised model follows all its settings.

Training Set	Testing Set	Market		Duke		MSMT	
	Supervision	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Market	MT [61]	*78.4*	*52.1*	–	–	–	–
	Weakly	*83.3*	*60.4*	48.9	29.7	**26.8**	**9.6**
	Fully	***91.3***	***77.3***	**58.7**	**38.2**	25.3	9.5
Duke	MT	–	–	*65.2*	*44.7*	–	–
	Weakly	68.4	37.7	*73.9*	*54.4*	33.7	11.9
	Fully	**72.7**	**43.0**	***82.5***	***67.3***	**35.4**	**13.0**
MSMT	MT	–	–	–	–	*39.6*	*15.9*
	Weakly	68.3	37.2	59.2	38.2	*49.4*	*21.5*
	Fully	**73.7**	**45.0**	**66.2**	**46.7**	***72.8***	***42.9***

ter used for learning the appearance features. We compare the performance of using all annotations (*fully-supervised learning*) and only intra-camera annotations (*weakly-supervised learning*). The results are in Tab. 3. For weakly-supervised experiments, we follow the same settings in MT [61]. Since there are no inter-camera annotations, the identity labels of different cameras are independent, and we assign each individual camera with a separate classifier. Each of these classifiers is supervised by the corresponding intra-camera identity labels. Surprisingly, even without inter-camera annotations, the *weakly-supervised learning* achieves competitive performance. According to these results, we believe that the importance of intra-camera annotations is significantly undervalued.

Transfer. In this section, we evaluate the ability to transfer ReID knowledge between the old and new datasets. First, we evaluate the ability to transfer previous knowledge to new domains. The related task is *domain adaptation*, which usually involves a labeled source training set and another unlabeled target training set. We integrate our formulation into a recent state-of-the-art method ECN [57]. The results are shown in Tab. 4. By aligning the distributions of source labeled images and target unlabeled images, the performance of ECN is largely boosted, *e.g.*, when transferring from Duke to Market, the Rank-1 accuracy and mAP are improved by 6.6% and 9.0%, respectively. Meanwhile, compared to other methods that also utilize camera labels, such as CamStyle [58] and CASCL [45], our method outperforms them significantly. These improvements demonstrate the effectiveness of our camera-based formulation in *domain adaptation*.

Second, we evaluate the ability to preserve old knowledge as well as accumulate common knowledge for all seen datasets when being fine-tuned. *Incremental learning*, which fine-tunes a model on a sequence of training sets, is used for this evaluation. Experiments are designed as follows. Given three large-scale ReID datasets, there are in total six training sequences of length 2, such as (Market→Duke) and six sequences of length 3, such as (Market→Duke→MSMT). We use the baseline method described in Sec. 3.4 and train it on all sequences separately. After training on each dataset of every sequence, we evaluate the

Table 4. The results of testing ReID models across datasets. ‡ marks methods that only use the source domain data for training, *i.e.*, direct transfer. Other methods listed in this table utilize both the source and target training data, *i.e.*, domain adaptation.

Method	Duke to Market				Market to Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
UMDL [27]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
PTGAN [42]	38.6	-	66.1	-	27.4	-	50.7	-
PUL [4]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
SPGAN [3]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
BoT*‡ [22]	53.3	69.7	76.4	24.9	43.9	58.8	64.9	26.1
MMFA [17]	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7
TJ-AIDL [41]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
CamStyle [58]	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
HHL [56]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
CASCL [45]	64.7	80.2	85.6	35.6	51.5	66.7	71.7	30.5
ECN [57]	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
Baseline‡	53.2	70.0	76.0	25.1	37.0	52.6	58.9	20.7
Ours+BoT*‡	68.6	82.5	87.7	39.0	60.6	74.0	78.5	39.8
Ours+Baseline‡	72.7	85.8	90.7	43.0	58.7	74.1	78.1	38.2
Ours+ECN	**81.7**	**91.9**	**94.7**	**52.0**	**68.0**	**80.0**	**83.9**	**44.9**

latest model on the first dataset of the corresponding sequence and record the performance decreases. Both the **Data-Free** and **Replay** settings are tested. For the Replay settings, the exemplars are selected by randomly sampling one image for each identity. Compared to the original training sets, the size of the exemplar set for Market, Duke, and MSMT is only 5.5%, 4.2%, and 3.4%, respectively. Note that in Replay settings, the old classifiers will also be updated in training. The corresponding results are shown in Tab. 5. To better demonstrate our improvements, we report the averaged results of the sequences that are of the same length and share the same initial dataset, *e.g.*, averaging the results of testing Market on the sequences Market→Duke and Market→MSMT. In short, our formulation outperforms the dataset-based formulation in all experiments. These results further demonstrate the effectiveness of our formulation.

4.3 Ablation Study

The experiments above demonstrate that our camera-based formulation boosts all the mentioned tasks. Now, we conduct more ablation studies to validate CBN.

Comparisons between CBN and other BN variants. We compare CBN with three types of BN variants. (1) BN [9] and IBN [26] correspond to the methods that use training-set-specific statistics to normalize all testing data. (2) AdaBN [15] is a dataset-wise adaptation that utilizes the testing-set-wise statistics to align the entire testing set. (3) The combination of BN and our CBN is to verify the importance of training ReID models with CBN. As shown

Table 5. Results of ReID models on incremental learning tasks. Each result denotes the percentage of the performance preserved on the first dataset after learning on new datasets. § marks the Data-Free settings. † corresponds to the Replay settings.

Testing Set		Market		Duke		MSMT	
Seq Length	Formulation	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1	–	100%	100%	100%	100%	100%	100%
2	Conventional§	82.2%	62.5%	80.2%	68.8%	55.5%	38.7%
	Ours§	88.3%	71.2%	89.3%	83.2%	74.5%	58.9%
	Conventional†	92.5%	84.1%	90.9%	84.7%	81.7%	70.1%
	Ours†	**95.0%**	**85.7%**	**94.3%**	**91.1%**	**91.6%**	**84.6%**
3	Conventional§	74.8%	52.2%	75.2%	63.0%	38.9%	24.7%
	Ours§	85.8%	66.0%	85.8%	77.4%	56.6%	39.4%
	Conventional†	86.5%	74.0%	84.1%	76.4%	74.3%	60.9%
	Ours†	**94.4%**	**83.1%**	**91.5%**	**87.6%**	**86.4%**	**76.0%**

Table 6. Results of combining different normalization strategies in fully-supervised learning and direct transfer. In this table, BN and IBN correspond to the training-set-specific normalization methods. AdaBN adapts the dataset-wise normalization statistics. CBN follows our camera-based formulation and aligns each camera independently.

Training Method	Testing Method	Duke to Duke		Duke to Market	
		Rank-1	mAP	Rank-1	mAP
BN	BN	81.5	66.6	53.2	25.1
IBN [26]	IBN	77.6	57.0	61.7	29.5
BN	AdaBN [15]	81.2	66.2	55.8	28.1
BN	Our CBN	80.2	63.7	69.5	40.6
Our CBN	Our CBN	**82.5**	**67.3**	**72.7**	**43.0**

in Tab. 6, training and testing the ReID model with CBN achieves the best performance in both *fully-supervised learning* and *direct transfer*.

Samples Required for CBN Approximation. We conduct experiments for approximating the camera-related statistics with different numbers of samples. Note that if a camera contains less than the required number of images, we simply use all available images rather than duplicate them. We repeat all experiments 10 times and list the averaged results in Tab. 7. As demonstrated, the performance is better and more stable when using more samples to estimate the camera-related statistics. Besides, results are already good enough when only utilizing very few samples, *e.g.*, 10 mini-batches. For the balance of simplicity and performance, we adopt 10 mini-batches for approximation in all experiments.

Compatibility with Different Backbones. Apart from ResNet [7] used in the above experiments, we further evaluate the compatibility of CBN. We embed CBN with other commonly used backbones: MobileNet V2 [30] and ShuffleNet V2 [23], and evaluate their performance on *fully-supervised learning* and *direct transfer*. As shown in Tab. 8, the performance is also boosted significantly.

Table 7. The mAP of our method on fully-supervised learning and direct transfer. We repeat each experiment 10 times and calculate the mean and variance of all results.

# Batches	Market to Market		Market to Duke	
	mean	variance	mean	variance
1	76.29	0.032	37.34	0.047
5	77.21	0.010	38.08	0.017
10	77.33	0.007	38.19	0.008
20	77.37	0.005	38.18	0.002
50	**77.39**	**0.001**	**38.21**	**0.001**

Table 8. Results of combining our camera-based formulation with different convolutional backbones. The fully-supervised learning results are in *italics*.

Backbone	Training Set	Testing Set	Market		Duke	
		Formulation	Rank-1	mAP	Rank-1	mAP
MobileNet V2 [30]	Market	Conventional	*87.7*	*69.2*	34.7	18.9
		Ours	***89.8***	***73.7***	**54.4**	**34.0**
	Duke	Conventional	51.4	22.6	*79.8*	*60.2*
		Ours	**70.7**	**39.0**	***79.9***	***62.4***
ShuffleNet V2 [23]	Market	Conventional	*82.6*	*58.4*	34.6	18.4
		Ours	***85.9***	***65.8***	**53.8**	**33.8**
	Duke	Conventional	48.1	20.3	*74.7*	*52.8*
		Ours	**70.0**	**38.9**	***77.1***	***58.6***

5 Conclusions

In this paper, we advocate for a novel camera-based formulation for person re-identification and present a simple yet effective solution named camera-based batch normalization. With only a few additional costs, our approach shrinks the gap between intra-camera learning and inter-camera learning. It significantly boosts the performance on multiple ReID tasks, regardless of the source of supervision, and whether the trained model is tested on the same dataset or transferred to another dataset. Our research delivers two key messages. **First**, it is crucial to align *all* camera-related distributions in ReID tasks, so the ReID models can enjoy better abilities to generalize across different scenarios as well as transfer across multiple datasets. **Second**, with the aligned distributions, we unleash the potential of intra-camera annotations, which may have been undervalued in the community. With promising performance under the weakly-supervised setting (only intra-camera annotations are available), our approach provides a practical solution for deploying ReID models in large-scale, real-world scenarios.

Acknowledgements

This work was supported by National Science Foundation of China under grant No. 61521002.

References

1. Almazan, J., Gajic, B., Murray, N., Larlus, D.: Re-id done right: towards good practices for person re-identification. arXiv preprint arXiv:1801.05339 (2018)
2. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: CVPR. IEEE (2018)
3. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR. IEEE (2018)
4. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **14**(4), 83 (2018)
5. Fan, X., Luo, H., Zhang, X., He, L., Zhang, C., Jiang, W.: Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In: ACCV. Springer (2018)
6. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: ICCV. IEEE (2019)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. IEEE (2016)
8. Huang, H., Yang, W., Chen, X., Zhao, X., Huang, K., Lin, J., Huang, G., Du, D.: Eanet: Enhancing alignment for cross-domain person re-identification. arXiv preprint arXiv:1812.11369 (2018)
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
10. Jiao, J., Zheng, W.S., Wu, A., Zhu, X., Gong, S.: Deep low-resolution person re-identification. In: AAAI (2018)
11. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: CVPR. IEEE (2018)
12. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
13. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR. IEEE (2018)
14. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR. IEEE (2018)
15. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. arXiv preprint arXiv:1603.04779 (2016)
16. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2018)
17. Lin, S., Li, H., Li, C.T., Kot, A.C.: Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In: BMVC (2018)
18. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: AAAI (2019)
19. Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: CVPR. IEEE (2020)
20. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing **26**(7), 3492–3506 (2017)

21. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: CVPR. IEEE (2018)
22. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPRW. IEEE (2019)
23. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: ECCV. Springer (2018)
24. Mao, S., Zhang, S., Yang, M.: Resolution-invariant person re-identification. arXiv preprint arXiv:1906.09748 (2019)
25. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: ICCV. IEEE (2019)
26. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV. Springer (2018)
27. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: CVPR. IEEE (2016)
28. Qian, N.: On the momentum term in gradient descent learning algorithms. Neural networks **12**(1), 145–151 (1999)
29. Rannen, A., Aljundi, R., Blaschko, M.B., Tuytelaars, T.: Encoder based lifelong learning. In: ICCV. IEEE (2017)
30. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. IEEE (2018)
31. Shen, Y., Li, H., Yi, S., Chen, D., Wang, X.: Person re-identification with deep similarity-guided graph neural network. In: ECCV. Springer (2018)
32. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: CVPR. IEEE (2018)
33. Song, J., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: CVPR. IEEE (2019)
34. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: Theory and practice. Pattern Recognition (2020)
35. Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K.: Part-aligned bilinear representations for person re-identification. In: ECCV. Springer (2018)
36. Sun, H., Chen, Z., Yan, S., Xu, L.: Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification. In: ICCV. IEEE (2019)
37. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: ICCV. IEEE (2017)
38. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV. Springer (2018)
39. Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J., Yan, J., Wang, X.: Eliminating background-bias for robust person re-identification. In: CVPR. IEEE (2018)
40. Van Der Maaten, L.: Accelerating t-sne using tree-based algorithms. JMLR **15**(1), 3221–3245 (2014)
41. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR. IEEE (2018)
42. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR. IEEE (2018)
43. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: Glad: Global-local-alignment descriptor for pedestrian retrieval. In: ACMMM. ACM (2017)

44. Wu, A., Zheng, W.S., Guo, X., Lai, J.H.: Distilled person re-identification: Towards a more scalable system. In: CVPR. IEEE (2019)
45. Wu, A., Zheng, W.S., Lai, J.H.: Unsupervised person re-identification by camera-aware similarity consistency learning. In: ICCV. IEEE (2019)
46. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV. IEEE (2017)
47. Yu, H.X., Wu, A., Zheng, W.S.: Unsupervised person re-identification by deep asymmetric metric embedding. TPAMI (2018)
48. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: CVPR (2019)
49. Zhang, T., Xie, L., Wei, L., Zhang, Y., Li, B., Tian, Q.: Single camera training for person re-identification. AAAI (2020)
50. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J.: Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
51. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV. IEEE (2015)
52. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications **14**(1), 13 (2017)
53. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV. IEEE (2017)
54. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR. IEEE (2017)
55. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)
56. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: ECCV. Springer (2018)
57. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: CVPR. IEEE (2019)
58. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: CVPR. IEEE (2018)
59. Zhou, J., Yu, P., Tang, W., Wu, Y.: Efficient online local metric adaptation via negative samples for person reidentification. In: ICCV. IEEE (2017)
60. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV. IEEE (2019)
61. Zhu, X., Zhu, X., Li, M., Murino, V., Gong, S.: Intra-camera supervised person re-identification: A new benchmark. In: ICCVW. IEEE (2019)
62. Zhu, Z., Jiang, X., Zheng, F., Guo, X., Huang, F., Sun, X., Zheng, W.: Viewpoint-aware loss with angular regularization for person re-identification. In: AAAI (2020)