Person30K: A Dual-Meta Generalization Network for Person Re-Identification

Yan Bai[1,4], Jile Jiao[2], Wang Ce[1], Jun Liu[3], Yihang Lou[1], Xuetao Feng[2], and Ling-Yu Duan[1,4,*]
[1]Peking University, Beijing, China [2]Alibaba Group, Beijing, China
[3]Singapore University of Technology and Design [4]Peng Cheng Laboratory, Shenzhen, China
{yanbai, wce, yihanglou, lingyu}@pku.edu.cn,
{jile.jjl, xuetao.fxt}@alibaba-inc.com, jun_liu@sutd.edu.sg

Abstract

Recently, person re-identification (ReID) has vastly benefited from the surging waves of data-driven methods. However, these methods are still not reliable enough for real-world deployments, due to the insufficient generalization capability of the models learned on existing benchmarks that have limitations in multiple aspects, including limited data scale, capture condition variations, and appearance diversities. To this end, we collect a new dataset named Person30K with the following distinct features: 1) a very large scale containing 1.38 million images of 30K identities, 2) a large capture system containing 6,497 cameras deployed at 89 different sites, 3) abundant sample diversities including varied backgrounds and diverse person poses. Furthermore, we propose a domain generalization ReID method, dual-meta generalization network (DMG-Net), to exploit the merits of meta-learning in both the training procedure and the metric space learning. Concretely, we design a "learning then generalization evaluation" meta-training procedure and a meta-discrimination loss to enhance model generalization and discrimination capabilities. Comprehensive experiments validate the effectiveness of our DMG-Net.

1. Introduction

Person re-identification (ReID) targets at matching people across different cameras. Recent years have witnessed remarkable progress of learning-based person ReID methods [5, 14, 20, 24, 34, 37], which have achieved promising performances when the training and testing sets are collected from the same scenarios or camera sets. However, they are often confronted with an inevitable accuracy drop when handling unseen cameras [3, 38]. This performance disparity reveals the problem of their limited generalization capability against data domain gaps, which are usually caused by different capture viewpoints, camera types, backgrounds and illumination conditions. Therefore, the study on enhancing ReID model's generalization ability is of great potential to explore for better practical deployment.

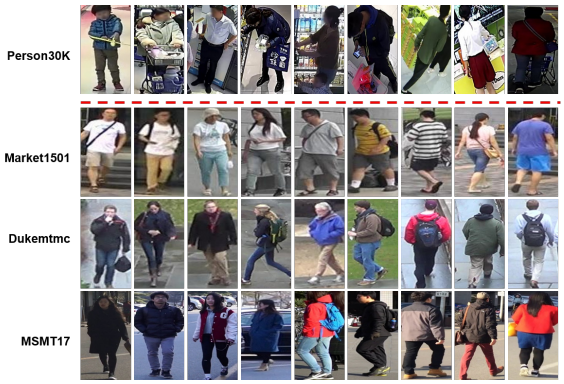

Figure 1. Comparison among samples from our Person30K, Matket1501 [47], Dukemtmc [28], and MSMT17 [38] datasets. Our Person30K covers more varied backgrounds, illumination conditions, person poses, and capture viewpoints.

Particularly, the benchmark dataset plays a crucial role in model training and comprehensive evaluation. However, the existing person ReID datasets still have many limitations, such as: (1) limited image samples and annotated identities (∼4K only); (2) very few capturing cameras (2∼15 only) involved in data collection; (3) less varied scenes and environmental conditions, such as backgrounds and illumination; (4) monotonous pedestrian poses, usually covering walking poses only. These limitations oversimplify the challenges of real-world person ReID, and restrict the representation and generalization capability of the models developed on these datasets.

To facilitate person ReID, we create a new dataset, named Person30K, with a very large scale and high data diversity. Our dataset has the following major advantages over existing benchmarks: (1) It contains 1.38 million images of 30K identities, 11 times the scale of the existing largest benchmark MSMT17 [38]. (2) The images in our Person30K are collected from 6,497 cameras deployed at 89

*Ling-Yu Duan is the corresponding author.

different sites, far exceeding the 15-camera setting used by MSMT17 [38]. (3) Person30K is captured both indoors and outdoors from multiple supermarkets and shopping malls, covering various scenes, *e.g.*, supermarket aisles, cashiers, mall corridors, restaurants, streets, and parking lots. (4) The person samples present very diversified postures and dressing styles. (5) We provide rich annotations for each sample, including person ID, camera ID, site ID, and scenario category (*e.g.*, supermarket and shopping mall). Moreover, our Person30K dataset can be used to evaluate model generalization at different domain gap levels with different testing subset divisions, according to varied training-testing data capture settings, such as the "same-site", "same-scenario", and "different-scenarios" configurations.

Based on the proposed Person30K dataset, in this paper, we focus on the domain generalizable person ReID. We aim to learn a model on a set of source domains/cameras, which can generalize well to other unseen domains/cameras without further fine-tuning. Unlike unsupervised domain adaptation methods [3, 38, 48] that still require using unlabeled target samples to adapt the model to target domain, the generalizable person ReID [13, 31, 49] is more convenient for practical deployment, for being free from the reliance on target domain samples. For example, for a large-scale ReID system deployed at chain supermarkets, generally it is unaffordable to go through the repetitive data re-collection, annotation, and model fine-tuning for every single supermarket. Therefore, we expect a generalizable ReID system to work out-of-the-box for widespread deployment.

For domain generalizable person ReID, we propose a novel Dual-Meta Generalization Network (DMG-Net), which is the first approach to exploit the meta-learning scheme in model training procedure and metric space learning simultaneously. Meta-learning, also referred to as learning to learn, aims at obtaining knowledge from "tasks/experiences". Such "tasks" generally mimic the target testing scenarios in model training process [4, 30]. We aim to exploit such a task construction and mimicking scheme in generalizable ReID scenario. Concretely, for our proposed meta-learning training procedure, we construct the tasks through a virtual "learning then generalization evaluation" process. We divide the training data into support sets (for virtual training) and query sets (for virtual evaluation) to mimic the domain generalization scenarios, where the support and query sets do not share any overlapping cameras. We train a base model on support sets and then evaluate its performance generalized on query sets. Thereby, based on the evaluation results/losses, we can update the model towards better generalization. Besides, to enhance model discrimination, we further propose a meta-discrimination loss to obtain a better metric space. We construct discrimination oriented meta tasks by explicitly mimicking the cross-camera sample matching process. Specifically, we compute feature-based centers to represent the identities in the support set, and conduct a matching process between the query samples and support centers. Overall, with the proposed dual meta-learning scheme, DMG-Net can strengthen model generalization on the unseen domain data, as well as model discrimination for cross-camera matching in person ReID.

Our major contributions are as follows. (1) We create a large-scale person ReID benchmark, Person30K, which contains 1.38 million images of 30K identities, collected from 6,497 cameras deployed at 85 supermarkets and 4 large shopping malls. We used one full year to collect Person30K and it cost us 130 man-months' labor to clean and annotate it. (2) We propose a DMG-Net for generalizable person ReID which exploits the meta-learning scheme in both the training procedure and metric space learning. A "learning then generalization evaluation" training procedure is designed to extend model generalization to unseen domains. And a discrimination loss is also incorporated for enhancing model discrimination in cross-camera matching.

2. Related Work

2.1. Person ReID Datasets

Benchmark datasets play a crucial role for model training and evaluation. Recently released person ReID datasets, including MSMT17 [38], Market1501 [47], Dukemtmc [28], and CUHK03 [18], are evolving towards larger scale than VIPeR [7], PRID [10] GRID [22] and i-LIDS [39]. Specifically, MSMT17 [38] contains the largest number of identities (∼4K) among these widely used datasets, as shown in Table 1. Though considerable person ReID accuracy has been achieved on these datasets, models trained on them are still not powerful enough due to the following limitations: 1) The limited data scale of existing ReID datasets, especially the limited sample numbers and person identities (1K∼4K only), understates the difficulty of the similar person retrieval problem in practice. 2) Generally, there are only 2∼15 cameras involved in their data collections, resulting in insufficiency of capture condition variations, *e.g.*, backgrounds and illuminations. 3) The person samples are mainly pedestrians, usually presenting monotonous walking poses, which limits model generalization potential towards handling samples with very different poses. 4) Existing datasets cover limited scenes captured by a single camera system. Owing to the limitations above, these datasets are not powerful enough for evaluating the representation and generalization capabilities of models.

2.2. Person ReID Methods

Supervised person ReID. For fully supervised person ReID, the training samples are all annotated with identity labels, and the training and testing data are captured by the

same camera system. Lots of efforts have been put into integrating effective mechanisms into supervised learning framework to extract more discriminative features for person ReID, including spatial alignment [34, 45], visual attention [19, 20], semantic segmentation [14, 37], and generative data augmentation [21, 24]. However, such full supervision methods often underperform when they are directly applied to new domains/camera sets [3, 5].

Domain adaptation for person ReID. Domain adaptation generally has a labeled source dataset and an unlabeled target dataset as the fundamental setup, and aims to obtain a discriminative model on the target domain. Adaptation based methods usually transfer the attributes or styles from labeled source data to unlabeled target domain to narrow down the domain gap [3, 11, 38, 48], or discover proper pseudo labels for target samples in training [5, 6, 32, 42]. However, it is overly tedious in application to treat each new camera/site as a new domain to conduct model adaptation, considering the difficulty of obtaining data from all domains beforehand and the cost of repetitive adaptation [13, 31]. Therefore, the domain generalization capability of ReID models starts to draw increasing attention.

Domain generalization for person ReID. Domain generalization aims to improve the generalization capability of the models trained on source data to unseen target domains, involving no target data in training. Such a new setup has been introduced to person ReID recently. Jin *et.al.* [13] proposed a style normalization and restitution module to distill identity-relevant features to bypass the domain gap issue. Zhuang *et.al.* [49] proposed a camera-based batch normalization to shrink the distribution gaps between different cameras. Besides, Song *et.al.* [31] proposed a domain-invariant mapping network to learn a mapping between a person sample and its identity classifier by a memory bank. Specifically, its sampling process followed a meta-learning pipeline by treating a subset of IDs as a learning task. Different from [31], we propose a novel dual-meta generalization network exploiting the merits of meta-learning in both the training procedure and metric space learning.

2.3. Meta-Learning

Meta-learning, also known as learning to learn, focuses on improving model proficiency by learning more experience from the process of mimicking the target testing scenarios. The well-received application of meta-learning is few-shot learning, which can be roughly categorized into optimization-based and metric-based approaches. Most of the optimization-based methods are model-agnostic [4], which focus on designing a training process to learn a good weight initialization for fast adaptation on a new task, such as MAML [4] and its variants [15, 26, 27]. The metric-based methods aim to learn a good representation in feature space, and thereby the model can be directly used for a new task without further adaptation [30, 36]. Recently, there also arise several methods using meta-learning for generic domain generalization [8, 16]. MLDG [16] first proposed a model-agnostic training strategy for domain generalization, but it cannot be directly applied to person ReID, because it assumed that the source and target domains shared the same label space and was designed for a small scale classification (7 classes only). Besides, MFR [8] used the MLDG training scheme in face recognition.

Different from these methods, we propose a dual-meta generalization method, which is the first approach to exploit the optimization-based training strategy and metric-based feature space learning within a unified ReID framework. By such a new attempt, the domain generalization and model discrimination can be improved at the same time.

3. Dataset

3.1. Description of Person30K Dataset

We collect a new dataset, named Person30K, containing 1.38 million images of 30K different identities. The dataset collection spanned across one full year, covering all seasons. We spent 130 man-months' labor in annotating such an enormous number of image samples out of 12,994-hour-long original video data ($\sim$ 1,403 million video frames). They were collected from 6,497 cameras of 89 capture systems deployed at different supermarkets and shopping malls. For privacy considerations, we blur the facial regions for all person samples. Compared to existing widely used person ReID datasets, such as the MSMT17 [38], Market1501 [47], and Dukemtmc [28], our proposed Person30K dataset considers and covers more challenging factors in practical application scenarios, as shown in Table 1 and Fig. 2. Below, we summarize the characteristics of our Person30K dataset.

- **Super large data scale:** The scale of our Person30K dataset is 11 times the scale of the existing largest benchmark MSMT17 [38] (1.38 Million v.s. 126K).
- **Very large camera system:** Person30K are collected from 6,497 cameras of various types deployed in 89 capture systems at different sites.
- **Various capturing conditions:** Our data collection covers various scenarios including indoor scenes like supermarket aisles, cashiers and mall corridors, and outdoor scenes like squares, streets, and parking lots.
- **Diversified person poses and dressing styles:** The Person30K dataset contains very diversified and challenging human poses, like pushing shopping charts and picking objects. Besides, Person30K also covers various dressing styles from summer shirts to winter coats.
- **Rich same-identity samples:** Averagely, each identity in our Person30K dataset has $\sim$46 samples, captured

Table 1. Comparison between Person30K and other person ReID datasets. In this table, the mark "($n\times$)" in the Person30K column means that the scale of data or capture setup of our Person30K is n times the scale of MSMT17 [38].

Dataset	**Person30K**	MSMT17 [38]	Duke [28]	Market [47]	CUHK03 [18]	CUHK02 [17]	VIPeR [7]	PRID [10]
Samples	**1,384,940** (11×)	126,441	36,411	32,668	28,192	7,267	1,264	1,134
Identities	**30,000** (7.3×)	4,101	1,812	1,501	1,467	1,816	632	934
Cameras	**6,497** (433×)	15	8	6	2	2	2	2
Capture Sites	**89** (89×)	1	1	1	1	1	1	1
Avg Number of Cameras Passed per Identity	**16.01** (3.4×)	4.67	2.67	4.42	2	2	2	2
Seasons	4	1	1	1	1	1	1	1
Scene	outdoor, indoor	outdoor, indoor	outdoor	outdoor	indoor	indoor	outdoor	outdoor

Figure 2. Person30K covers various scenes captured both indoors and outdoors, and contains many challenging factors for ReID.

under 16 different cameras, which facilitates constructing a challenging cross-camera ReID setup.

- **Additional annotations:** Other than the identity labels, Person30K dataset also provides annotations, including camera labels, data capture sites and site scenarios (*e.g.*, supermarket and shopping mall).
- **Versatile dataset splitting:** The testing set of Person30K can be divided into different subsets according to the capturing scenarios, sites, and cameras, for comprehensive evaluation of model discrimination and generalization.

3.2. Evaluation Protocol

We divide Person30K into training and testing sets based on the capturing scenarios. The training set only includes the data captured from supermarket scenario, while the testing set covers both supermarket and mall scenarios, as shown in Table 2. Furthermore, we construct 3 different testing subsets with different characteristics:

- Test-A: "same-site" subset shares the same capturing cameras and sites as the training set, similar to the setting in existing datasets [28, 38, 47].
- Test-B: "same-scenario" subset covers different capturing cameras and sites compared to the training set, but under the same supermarket scenario.

Table 2. The splitting of Person30K's training and testing sets.

Set Scale	Train	Test-A	Test-B	Test-C
Identities	12,000	6,000	6,000	6,000
Images	568,977	287,876	268,743	259,344
Sites	60	60	25	4
Cameras	3,680	3,699	1,731	1,017
Avg Number of Cameras Passed by per Identity	9.18	13.68	13.31	34.70
Camera overlapping with training set	-	Yes	No	No
Captured scenarios	Markets	Markets	Markets	Malls

Table 3. The results of model direct transfer across datasets (mAP/CMC@1).

Source \ **Target**	Person30K	Market1501	Dukemtmc	MSMT17
Person30K	-	**71.1 / 86.3**	**57.3 / 72.2**	**32.9 / 57.2**
Market1501	4.1 / 17.7	-	28.9 / 44.5	7.0 / 19.4
Dukemtmc	3.4 / 15.95	31.4 / 60.4	-	9.8 / 28.1
MSMT17	9.0 / 33.4	34.5 / 64.0	42.3 / 62.6	-

- Test-C: "different-scenario" subset differs from the training set in terms of the capturing scenario, *i.e.* the mall scenario v.s. the supermarket scenario.

Under such testing set divisions, we can comprehensively evaluate the model generalization ability at different domain gap levels. As for the evaluation metrics, we adopt the widely used Cumulative Match Characteristics (CMC) and mean Average Precision (mAP) [38, 47].

3.3. Analysis on Dataset Diversity

To demonstrate the diversity of Person30K and its potential to facilitate model generalization, we conduct direct transfer experiments based on the ResNet50 backbone with softmax and triplet loss, *i.e.*, training a model on one dataset, and then testing it on other datasets. We adopt Market1501 [47], Dukemtmc [28], and MSMT17 [38] for the transfer comparison with Person30K. As shown in Table 3, the model trained on our Person30K can achieve 71.1%, 57.3%, 32.9% mAP on the Market1501, Dukemtmc, and MSMT17 datasets, while the model trained on these datasets can only achieve 4.1%, 3.4%, 9.0% mAP on Person30K. Besides, when we adopt Market1501 as the target dataset, the model trained on our Person30K can achieve 71.1% mAP largely outperforming the Dukemtmc and MSMT17 models (31.4% and 34.5% only). The above

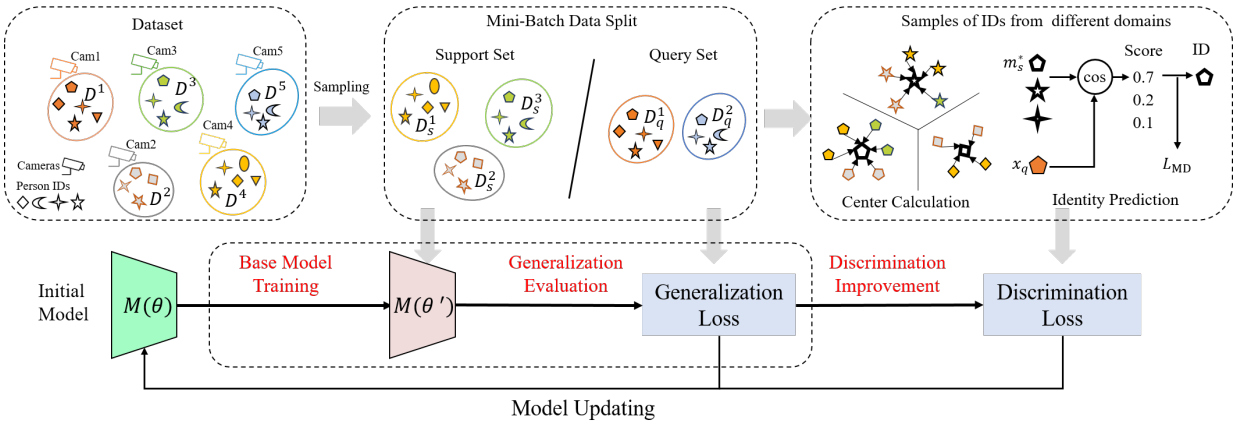

Figure 3. Illustration of our proposed DMG-Net. DMG-Net includes a meta-generalization training procedure and a meta-discrimination loss. The meta-generalization first trains a base model on the support set, then performs generalization evaluation on the query set. For meta-discrimination loss, it optimizes the metric space to improve model discrimination.

results indicate the strong potential of Person30K for promoting the research of Person ReID.

4. Proposed Methods

In this paper, we study a generalizable person ReID problem, where at the training stage, a model $M(\theta)$ is trained on a set of source cameras, and at the testing stage, the model is able to generalize well to a set of new unseen cameras without any model updating. To this end, we propose a meta-learning based Dual-Meta Generalization Network (DMG-Net). It leverages the benefits of meta-learning on extending model learning capabilities through bunches of "tasks" that mimic target testing scenarios. The term "dual" here indicates that our DMG-Net exploits the meta-learning merits from two perspectives, namely, a meta-generalization training procedure to achieve "learning to generalize", and a meta-discrimination constraint to achieve "learning to discriminate", as illustrated in Fig. 3.

Specifically, for the meta-generalization training procedure, we mimic the domain generalization scenario, where the training and testing samples are captured from non-overlapping cameras. For the meta-discrimination loss, we explicitly mimic the cross-camera matching process to optimize the metric space. Particularly, we accommodate the support-query data division in meta-learning pipelines to generalizable person ReID. Concretely, we divide the training images into support set $D_s = \{D_s^1, D_s^2, ..., D_s^{C/2}\}$, and query set $D_q = \{D_q^{C/2+1}, D_q^{C/2+2}, ..., D_q^C\}$, where C is the camera number, $D_*^c = \{(x_i, y_i, \mathrm{cam}_i)\}$ is the sample set from the c-th camera (*i.e.*, $\mathrm{cam}_i = c$), and * denotes s or q. Such a division can meet the requirements for both the meta-generalization and meta-discrimination.

4.1. Meta Generalization Network

To enable the ReID model to obtain the "learning to generalize" capability, we implement a "learning then generalization evaluation" training procedure based on meta-learning scheme. For generalizable ReID, we conduct tasks to mimic practical training and deploying scenarios, between which there are no overlapping cameras. To construct tasks, we split the training data into the support set D_s and query set D_q in mini-batches. We first use the support set to train the base model, and then use the query set to evaluate its generalization performance under the "deploying scenario" during each task optimization. Then, based on the evaluation results (losses), we further update the model towards better generalization. This two-stage "learning then generalization evaluation" optimization procedure is shown in Algorithm 1.

Base model training. Based on the support set, the base model training can be optimized like other supervised methods. Here, the training loss $\mathcal{L}_B$ of base model consists of the widely used softmax loss $\mathcal{L}_{\mathrm{soft}}$ and triplet loss $\mathcal{L}_{\mathrm{tri}}$ [23, 43] as follows,

$$\mathcal{L}_B(\theta) = \mathcal{L}_{\mathrm{soft}}(D_s; \theta) + \mathcal{L}_{\mathrm{tri}}(D_s; \theta), \tag{1}$$

where θ is the initial parameters of model $M(\theta)$ before base model training. Then, based on $\mathcal{L}_B$, we can obtain the base model parameters $\theta^{'}$ by a standard gradient update,

$$\theta^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_B(\theta). \tag{2}$$

where α is the learning rate hyper parameter.

Model generalization evaluation. After obtaining the updated $\theta^{'}$ on the support set that is sampled from several cameras, we evaluate its generalization capability under the query set sampled from other cameras. Such an evaluation on the query set is performed on the base model $M(\theta^{'})$, with the generalization loss $\mathcal{L}_G$ as follows,

$$\mathcal{L}_G(\theta^{'}) = \mathcal{L}_{\mathrm{soft}}(D_q; \theta^{'}) + \mathcal{L}_{\mathrm{tri}}(D_q; \theta^{'}). \tag{3}$$

Such an evaluation simulates the testing on unseen cameras, so as to make the model learn to generalize.

Algorithm 1 Meta-optimization procedure

Input: (1). A pre-trained model $M(\theta)$ parametrized by θ. (2). Training dataset D, which can be organized as $D = \{D_1, D_2, ..., D_C\}$, where C is the number of cameras, and $D_c = \{(x_i, id_i, cam_i)\}_{i=1}^{N_c}$ is the set containing N_c samples with camera ID c (*i.e.*, cam_i = c).

for total training epochs **do**
  Randomly select the non-overlapping support camera set and query camera set.
  Obtain the support set D_s and query set D_q according to the camera division.
  for k batches **do**
    Sample a mini-batch of support samples.
    Calculate loss $\mathcal{L}_B(\theta)$ using support images.
    Compute adapted parameters $\theta^{'} = \theta - \alpha\nabla_\theta\mathcal{L}_B(\theta)$.
    Sample a mini-batch of query samples.
    Calculate loss $\mathcal{L}_G(\theta^{'})$ using query samples.
    Calculate the gradient $\nabla_\theta L_G(\theta^{'})$.
    Update $\theta \leftarrow \theta - \alpha\nabla_\theta L_G\theta^{'}$.
  end for
end for

Meta-optimization. The whole objective thus becomes minimizing the $L_G(\theta^{'})$ for optimized parameters concerning generalization,

$$\min_{\theta} \mathcal{L}_G(\theta^{'}) = \min_{\theta} \mathcal{L}_G(\theta - \alpha\nabla_\theta\mathcal{L}_B(\theta)). \tag{4}$$

Note that, the meta-optimization is performed over the initial model parameters θ, and $\theta^{'}$ is only an intermediate result used to evaluate the model generalization. Such a meta-optimization is inspired by MAML [4]. However, different from MAML that focuses on learning good initialization parameters for a network, and needs additional adaptation training to a new task, we aim to obtain a model with strong generalization capability without any model updating for handling new cameras. More importantly, for MAML, to obtain a good model initialization across all tasks, it concurrently learns multiple tasks first, then updates the model by the mean gradient of these tasks. Different from MAML and its variants [4, 8, 16], we can obtain generalization ability within one mimicking task by learning base model on support set and generalizing to query set.

4.2. Meta Discrimination Loss

Apart from the aforementioned generalization problem that requires models to generalize well to unseen domain data, the discrimination of identities across different cameras is also a core challenge in person ReID. Therefore, we further explore a meta discrimination loss to realize "learning to discriminate" in a metric-based meta-learning fashion [30]. Our designed meta discrimination task explicitly mimics the cross-camera matching process in metric loss computation, which aims to enforce the samples of the same ID but captured from different cameras to get closer for better discrimination against data source variations.

Specifically, in a mini-batch, the matching process is conducted between the support set D_s and query set D_q, and D_s and D_q here are sampled from the same P identities, yet from different cameras. To implement such a cross-camera matching, we design a novel P-way classification, which classifies a query sample to P support IDs within a mini-batch. Particularly, this P-way classification acts as the "meta task" for our meta discrimination loss. To represent an ID p in support set for classification, we compute the mean feature m_s^p of the samples belonging to the ID p as the class center,

$$m_s^p = \frac{1}{N_p} \sum_{\substack{x_s \in D_s \\ \text{and } y_s = p}} f_\theta(x_s), \tag{5}$$

where N_p is the sample number of person p in support set in a mini-batch, and y_s is the label / ID of support sample x_s. $f_\theta(.)$ means the feature extracted by $M(\theta)$. Thereby for the P identities in a mini-batch, we can obtain P support centers $\{m_s^p\}_{p=1}^P$ to represent all of them. With such support centers and the query sample, we can explicitly conduct the matching process by a "P-way classification".

Note that since the sampled P IDs in our meta task will change in each mini-batch, the fully connected layer based P-way probability prediction for fixed classes is not applicable. Thus we adopt a similarity-based strategy. We first compute the cosine similarities between query feature $f_\theta(x_q)$ and support centers $\{m_s^p\}_{p=1}^P$, and then predict which support IDs the query belongs to based on the similarity scores. Therefore, the constraint of meta discrimination loss $\mathcal{L}_{\text{MD}}$ can be formulated as,

$$\mathcal{L}_{\text{MD}} = -\log \frac{\exp(\langle f_\theta(x_q), m_s^p \rangle)}{\sum_{p'=1}^{P} \exp(\langle f_\theta(x_q), m_s^{p'} \rangle)}, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. $f_\theta(x_q)$ and m_s^p belong to the same identity p. Both $f_\theta(x_q)$ and m_s^p are L2-normalized. By such design, $\mathcal{L}_{\text{MD}}$ is able to maximize the similarity between $f_\theta(x_q)$ and m_s^p, meanwhile minimizing the similarities to all other support centers (*i.e.*, towards 0).

The proposed $\mathcal{L}_{\text{MD}}$ enjoys three major advantages for the ReID task. First, by matching the query samples to the support centers, we explicitly construct the cross-camera matching process, which enables learning a discriminative model. Second, $\mathcal{L}_{\text{MD}}$ optimizes the relationships between query samples to all support samples/centers in a mini-batch, which is more effective than triplet loss and Center loss [40] only optimizing the relationships among three samples or the same ID samples. Third, we obtain the prediction probability by cosine similarities instead of designing a P-way classifier (*e.g.*, fully connected layer) [4, 31], which thus avoids repetitively updating classifiers when sampling new identities in mini-batches, so we can focus on learning a more discriminative and stable metric space.

Table 4. Performance comparison on the Person30K dataset.

Settings	Test-A			Test-B			Test-C			
Methods	mAP	CMC@1	CMC@5	mAP	CMC@1	CMC@5	mAP	CMC@1	CMC@5	Year
Softmax	65.72	79.40	91.05	57.63	80.15	91.23	57.32	83.83	95.30	-
CBN [49]	66.51	81.43	91.83	63.18	83.40	93.72	62.57	85.85	96.03	2020
DFLNet [1]	69.50	82.81	92.43	62.12	81.35	92.88	62.04	85.76	96.47	2020
Center Loss [40]	69.24	83.02	92.70	61.32	80.45	92.15	61.87	85.70	96.93	2016
Circle Loss [33]	70.91	84.03	93.13	62.05	82.73	94.02	61.96	84.23	96.45	2020
AGW [44]	71.04	83.57	93.70	63.12	82.12	93.37	63.12	85.85	96.03	2020
BoT [23]	69.19	80.55	90.32	62.81	77.61	91.97	62.01	84.73	96.60	2019
DMG-Net	**72.19**	**84.23**	**93.95**	**64.44**	**83.60**	**94.08**	**64.39**	**86.08**	**96.78**	Ours

4.3. Overall Objective

Finally, to simultaneously optimize the ReID model for both better generalization and discrimination performances, we obtain the overall loss for our DMG-Net as,

$$\mathcal{L}_{\text{All}} = \mathcal{L}_G + \mathcal{L}_{\text{MD}}. \tag{7}$$

By such a design, the proposed dual-meta generalization network can strengthen model generalization on the unseen domain data in training set and model discrimination for cross-camera matching in person ReID at the same time.

5. Experiments

Dataset Settings. To comprehensively evaluate the discrimination and generalization capabilities of ReID models, we experiment on the proposed Person30K dataset, and two mixed multi-dataset benchmarks following previous works [13, 31]. Specifically, we denote the mixed benchmark in [31] as **Mixed A**, where CUHK02, CUHK03, Market1501, Dukemtmc, and CUHK-SYSU PersonSearch [41] are mixed up for training, and VIPeR, PRID, GRID, and i-LIDS are adopted as the test sets. The other mixed benchmark in [13] is denoted as **Mixed B**, where the training set includes MSMT17, CUHK03, Market1501, and Dukemtmc, while VIPeR, PRID, GRID, and i-LIDS are used for testing, same as the test sets in Mixed A.

Implementation Details. We adopt ResNet50 [9] as our backbone, employing the bag-of-tricks (BoT) strong baseline [23] scheme. The input image size is 256×128. The mini-batch contains 128 images in total, 64 for both support and query sets (4 images per ID $\times$ 16 IDs), by following the widely used hyper parameters as [23, 40]. The support and query sets are sampled from different cameras, yet with same identities. For data augmentation, we perform color jitter scheme on the Mixed benchmark. We optimize the model with Adam optimizer. The ReID model is trained for 120 epochs with the start learning rate of 3.5×10^{-4} and performs learning rate decay $1/10$ in the $40th$ and $70th$ epochs.

5.1. Evaluation on Person30K Dataset

Table 4 shows the comparison results on our Person30K dataset, including the widely used baseline method Bag-of-tricks(BoT) [23], classic Center loss [40], as well as the latest works AGW [44], Circle loss [33], CBN [49], DFLNet [1], and our DMG-Net. The Bag-of-tricks includes label smoothing, warmup strategy, data augmentation, and BNNeck schemes, which provides a baseline in our DMG-Net. Compared with these methods, our proposed DMG-Net achieves the best performance on all subset divisions. Particularly, on Test-A, DMG-Net also takes the first place, since our proposed meta-discrimination loss can enhance the model's cross-camera matching ability. Even compared with the latest AGW [44] and DFLNet [1], which improve model discrimination by attention scheme or disentangle scheme, we can also achieve superior performance. Besides, compared to the domain generalization method, CBN [49], which uses batch normalization to align the distribution of different camera data, DMG-Net also achieves better performance by the "learning then generalization evaluation" training procedure. The obvious advantages of DMG-Net on Test-B and Test-C, which have larger domain gaps between training and testing sets, validate the enhanced model generalization capability of our DMG-Net.

5.2. Evaluation on Mixed Datasets

Table 5 shows the domain generalization performances on four target datasets. We make comparisons among our DMG-Net, generic meta-learning methods PPA [26] and Reptile [25], generic domain generalization methods MLDG [16] and CrossGrad [29], domain aggregation models AGG_PCB [35] and AGG_Align [46], as well as ReID-specific generalization methods SNR [13], DIMN [31], DDAN [2] and DualNorm [12]. We observe that the generic methods like Reptile [25] and MLDG [16], which are variants of classic meta-learning work MAML, can only obtain 26.90% and 35.36% mAP on PRID in Mixed A setup. This is because they are designed for category-level recognition task, which limits their adaptation to person ReID. Differently, our proposed DMG-Net optimizes a uniform feature space between different tasks, and can achieve much better performance. As for ReID-specific methods, our DMG-Net beats the DIMN [31] by a remarkable margin of 16.43% mAP on PRID dataset. DIMN [31] aims to learn a mapping network by predicting classifier weight for each identity, while we focus on learning an embedding feature space

Table 5. Performance comparison on the mixed dataset.

Dataset	Method	PRID		GRID		VIPeR		iLIDs	
		mAP	CMC@1	mAP	CMC@1	mAP	CMC@1	mAP	CMC@1
Mixed A	PPA [26]	45.26	31.90	37.98	26.88	54.46	45.06	72.73	64.50
	Reptile [25]	26.90	17.90	23.02	16.24	31.33	22.06	67.11	56.00
	Agg_PCB [35]	32.04	21.50	44.66	36.00	45.38	38.10	73.92	66.67
	Agg_Align [46]	25.50	17.20	24.67	15.92	52.94	42.78	74.69	63.83
	MLDG [16]	35.36	24.00	23.57	15.76	33.52	23.51	65.18	53.83
	CrossGrad [29]	28.18	18.80	16.00	8.96	30.40	20.89	61.29	49.67
	DIMN [31]	51.95	39.20	41.09	29.28	60.12	51.23	78.39	70.17
	DualNorm [12]	64.90	60.40	45.70	41.40	58.00	53.90	78.50	74.80
	DDAN [2]	58.90	54.50	55.70	50.60	56.40	52.30	81.50	78.50
	BoT [23]	61.25	51.40	49.62	40.48	56.66	48.20	81.27	74.67
	DMG-Net(ours)	**68.38**	**60.60**	**56.62**	**50.96**	**60.38**	**53.91**	**83.94**	**79.33**
Mixed B	SNR [13]	60.00	49.00	41.30	30.40	65.00	55.10	**91.90**	**87.00**
	BoT [23]	59.12	48.50	38.72	29.52	66.34	60.06	85.57	80.83
	DMG-Net(ours)	**69.73**	**59.70**	**47.18**	**37.28**	**70.93**	**62.34**	88.20	83.00

Table 6. Ablation study.

Method	Test-A	Test-B	Test-C
baseline(Bag-of-Tricks)	69.19	62.81	62.01
Center Loss [40]	69.24	61.32	61.87
Circle Loss [33]	70.91	62.05	61.96
Meta-discrimination	71.01	62.49	62.47
Meta-generalization	70.07	63.72	63.97
DMG-Net	72.19	64.44	64.39

instead of the classifiers to obtain more generalizable and discriminative features. Compared to the state-of-the-art SNR [13] and DDAN [2] methods, which design style normalization module or domain alignment scheme to obtain invariant features, DMG-Net also achieves superior performances on most datasets, showing better model generalization potential.

5.3. Ablation Study

DMG-Net contains two prominent technical proposals: the meta-generalization training scheme and the meta-discrimination loss. As shown in Table 6, when incorporating loss-enhanced methods into the same baseline [23], our meta-discrimination loss can provide better performance than other methods, such as Center Loss [40] and Circle Loss [33]. It indicates that the explicit mimicking of the cross-camera matching process can improve model discrimination effectively. Besides, when the meta-generalization training procedure is further integrated, the model can achieve $\sim$2% mAP gains on Person30K subsets Test-B and Test-C, which validates the effectiveness of our dual-meta generalization network.

5.4. Result Visualization

Fig. 4 visualizes several retrieval results of our DMG-Net. In Fig. 4 (a), we can observe that given the query from an outdoor scene, DMG-Net can correctly recall ground-truth samples from indoor scenes despite the background and camera viewpoint variations. Even if there are occlusion and multiple persons in the queries, DMG-Net can still ignore those distractions and retrieve the same-identity images, as shown in Fig. 4 (b)(c). Additionally, the individual cases of false positives also present reasonable visual similarities to the query images. Such impressive retrieval results benefit from both the large-scale training dataset and the effective DMG-Net method.

Figure 4. Visualization of the DMG-Net Top10 ReID results. The black, green, and red boxes indicate the queries, correct and wrong results, respectively.

6. Conclusion

In this paper, we create a large-scale Person30K dataset that presents diversified data capturing conditions, which is expected to promote the research and deployment of ReID models in real-world scenarios. Besides, we propose a domain generalizable ReID method integrating meta-learning scheme into the model training procedure and metric space learning, to improve model generalization and discrimination capability. Extensive experiments demonstrate the effectiveness of the proposed method.

Acknowledgement: This work was supported by the National Natural Science Foundation of China under Grant 62088102, and in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation. This work was also supported in part by Alibaba Group through Alibaba Research Intern Program.

References

[1] Yan Bai, Yihang Lou, Yongxing Dai, Jun Liu, Ziqian Chen, and Ling-Yu Duan. Disentangled feature learning network for vehicle re-identification. *IJCAI*, 2020. 7

[2] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Qi Tian, and Rongrong Ji. Dual distribution alignment network for generalizable person re-identification. *arXiv preprint arXiv:2007.13249*, 2020. 7, 8

[3] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 994–1003, 2018. 1, 2, 3

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017. 2, 3, 6

[5] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 6112–6121, 2019. 1, 3

[6] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 3

[7] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 262–275. Springer, 2008. 2, 4

[8] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6172, 2020. 3, 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, pages 770–778, 2016. 7

[10] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proceedings of the Scandinavian conference on Image analysis*, pages 91–102, 2011. 2, 4

[11] Houjing Huang, Wenjie Yang, Xiaotang Chen, Xin Zhao, Kaiqi Huang, Jinbin Lin, Guan Huang, and Dalong Du. Eanet: Enhancing alignment for cross-domain person re-identification. *arXiv preprint arXiv:1812.11369*, 2018. 3

[12] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification:generalizing person re-id in practice. *arXiv preprint arXiv:1905.03422*, 2019. 7, 8

[13] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3143–3152, 2020. 2, 3, 7, 8

[14] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3

[15] Kwonjoon Lee and et al. Meta-learning with differentiable convex optimization. In *CVPR*, 2019. 3

[16] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. *arXiv preprint arXiv:1710.03463*, 2017. 3, 6, 7, 8

[17] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3594–3601, 2013. 4

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159, 2014. 2, 4

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018. 3

[20] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017. 1, 3

[21] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018. 3

[22] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010. 2

[23] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 2019. 5, 7, 8

[24] Shunan Mao, Shiliang Zhang, and Ming Yang. Resolution-invariant person re-identification. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 883–889, 2019. 1, 3

[25] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 7, 8

[26] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7229–7238, 2018. 3, 7, 8

[27] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, pages 113–124, 2019. 3

[28] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 17–35. Springer, 2016. 1, 2, 3, 4

[29] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 7, 8

[30] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 2, 3, 6

[31] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 719–728, 2019. 2, 3, 6, 7, 8

[32] 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*, page 107173, 2020. 3

[33] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. *arXiv preprint arXiv:2002.10857*, 2020. 7, 8

[34] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 3

[35] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 7, 8

[36] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1208, 2018. 3

[37] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3

[38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 79–88, 2018. 1, 2, 3, 4

[39] Zheng Wei-Shi, Gong Shaogang, and Xiang Tao. Associating groups of people. In *Proceedings of the British Machine Vision Conference*, pages 23–1, 2009. 2

[40] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2016. 6, 7, 8

[41] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2(2), 2016. 7

[42] 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. *arXiv preprint arXiv:1912.01349*, 2019. 3

[43] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020. 5

[44] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020. 7

[45] Xuan Zhang, Hao Luo, X. Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, W. Jiang, C. Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *ArXiv*, abs/1711.08184, 2017. 3

[46] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 7, 8

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 1, 2, 3, 4

[48] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188, 2018. 2, 3

[49] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 7