META: Mimicking Embedding via oThers' Aggregation for Generalizable Person Re-identification

Boqiang Xu[1,2] Jian Liang[1,2] Lingxiao He[3] Zhenan Sun[1,2*]

[1] CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences (CASIA)

[2] University of Chinese Academy of Sciences (UCAS) [3] JD AI Reasearch

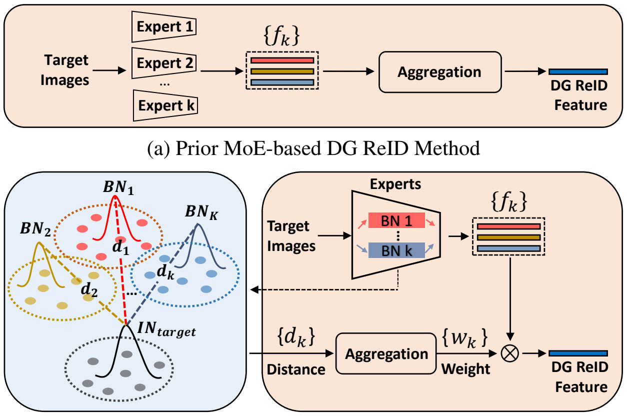
boqiang.xu@cripac.ia.ac.cn liangjian92@gmail.com helingxiao3@jd.com znsun@nlpr.ia.ac.cn

Abstract

Domain generalizable (DG) person re-identification (ReID) aims to test across unseen domains without access to the target domain data at training time, which is a realistic but challenging problem. In contrast to methods assuming an identical model for different domains, Mixture of Experts (MoE) exploits multiple domain-specific networks for leveraging complementary information between domains, obtaining impressive results. However, prior MoE-based DG ReID methods suffer from a large model size with the increase of the number of source domains, and most of them overlook the exploitation of domain-invariant characteristics. To handle the two issues above, this paper presents a new approach called Mimicking Embedding via oThers' Aggregation (META) for DG ReID. To avoid the large model size, experts in META do not add a branch network for each source domain but share all the parameters except for the batch normalization layers. Besides multiple experts, META leverages Instance Normalization (IN) and introduces it into a global branch to pursue invariant features across domains. Meanwhile, META considers the relevance of an unseen target sample and source domains via normalization statistics and develops an aggregation network to adaptively integrate multiple experts for mimicking unseen target domain. Benefiting from a proposed consistency loss and an episodic training algorithm, we can expect META to mimic embedding for a truly unseen target domain. Extensive experiments verify that META surpasses state-of-the-art DG ReID methods by a large margin.

1. Introduction

Person re-identification (ReID) aims at retrieving persons of the same identity across non-overlapping cameras. Many prior works [22, 25, 39, 44] have been devoted to the fully-supervised ReID task. Despite the promising performance when training and testing on the same domain, the performance always drops significantly when testing on an unseen domain because of the domain shift [41]. To avoid this, recent efforts are devoted to domain adaptive (DA) ReID [8,46,52] and domain generalizable (DG) ReID [5, 6, 14, 48]. In contrast to DA ReID, DG ReID is more practical and challenging as it utilizes training data from multiple source domains and directly tests across different and unseen domains, without any target data for training or fine-tuning. In this paper, we mainly focus on the challenging DG ReID problem.



(a) Prior MoE-based DG ReID Method

(b) Our Method

Figure 1. Differences between prior MoE-based DG ReID method and our method. (a) Prior MoE-based DG ReID methods add an individual network (expert) for each source domain, suffering from a large model size with the increase of the number of source domains. (b) Experts in our method share all the parameters except for the batch normalization layers. In the testing stage, we calculate the distance between IN statistics of test samples and the BN statistics of source domains for measuring the relevance of target domain w.r.t. source domains. Such distance is exploited by an aggregation network to adaptively integrate multiple experts.

Most of the prior DG ReID methods [1, 5, 14, 32, 48] assume an identical model for different domains. However,

[*] Corresponding author

such an assumption learns a common feature space for different source domains, which may neglect the individual domains' discriminative information and ignore the relevance of the target domain w.r.t source domains. To handle the issues above, mixture of experts (MoE) [13] has been studied for DG ReID, as shown in Fig. 1a. MoE can improve the generalization of models by integrating multiple domain-specific expert networks with the target domain's inherent relevance w.r.t. diverse source domains. Generally, prior MoE-based DG ReID methods have two potential problems: 1) As each source domain contains an individual branch network, the model size becomes fairly large with the increase of the number of source domains, limiting the practical deployment. 2) Most prior MoE-based DG ReID methods merely focus on learning domain-specific representations but overlook the domain-invariant characteristics.

To tackle the two issues above, we propose a novel DG ReID approach called Mimicking Embedding via Others' Aggregation (META), as shown in Fig. 1b. Batch Normalization (BN) statistics are able to imply the characteristics of the individual domain [28]. Inspired by this, instead of adding a branch network for each source domain, we train the META as a lightweight ensemble of multiple experts sharing all the parameters except for the domain-specific BN layers (*i.e*., one for each source domain for collecting domain-specific BN statistics). By doing so, META is able to exploit the diversified characteristics of each source domain and meanwhile, keeping the model size from increasing as the source domain increases. To extract the domain-invariant features, we design a global branch and leverage Instance Normalization (IN) [7], which works as a style normalization layer for filtering out domain-specific contrast information, to explicitly extract domain-invariant features.

Specifically, in our META method, we exploit individual domains' discriminative information by domain-specific BN layers. Then, during testing, the characteristics of the test samples from the unseen domain can be indicated by the means of their IN statistics. By measuring the distance between the IN statistics of the test samples and the BN statistics of source domains, we can infer the relevance of the unseen target domain w.r.t. source domains. Taking the relevance as input, we further devise a small aggregation network to integrate multiple experts for obtaining the accurate representation of the target person from an unknown domain. By doing so, those relevant source domains are able to contribute more valuable information than those less relevant domains. Moreover, we adopt episodic training [16] which simulates the test process at training time for updating the aggregation network. For each training batch, we collect training samples from the same source domain ($e.g.$, D_k) to simulate the 'unseen target data' for other domain experts. We propose a consistency loss to push the aggregated features of other domain experts as discriminative as the features extracted by the expert of D_K. In this way, the aggregation network is learned to be able to adaptively integrate diverse domain experts for explicitly mimicking any unseen target domain.

Our major contributions can be summarized as follows:

- We propose a novel META method to handle the DG ReID problem. Specifically, META leverages the domain-specific BN layers and designs a global branch to respectively tackle the two issues (*i.e*., model scalability and oversight in domain invariance) in prior MoE-based DG ReID methods.
- We develop a learnable aggregation network, updated by a proposed consistency loss and an episodic training algorithm, to adaptively integrate diverse domain experts via normalization statistics for mimicking any unseen target domain.
- Extensive experiments demonstrate that META surpasses state-of-the-art DG ReID methods by a large margin under various protocols.

2. Related Work

Domain Generalizable Person Re-identification. Person ReID has made great progress in recent years. Many methods have been proposed to improve the ReID performance, including but not limited to part-based methods [36,40], pose-based methods [20,26,34], and attention-based methods [45,47]. Despite the promising performance brought by these methods when training and testing on the same domain, the performance always drops significantly when testing on an unseen domain because of the domain shift [41]. To tackle this problem, some researchers start to study the unsupervised domain adaption (UDA) methods [8,46,52]. However, UDA requires unlabeled data from the target source, which is sometimes difficult to be collected in practical applications. As a result, domain generalizable (DG) ReID [5,6,14,48] have captivated researchers recently. Generally, DG ReID utilizes training data from multiple source domains and directly tests across different and unseen domains, without any target data for training or fine-tuning.

We briefly classify prior DG ReID methods into three categories. The first category is ***Meta-Learning*** [1, 5, 32, 48]. Meta-learning is a training strategy, which adopts the concept of 'learning to learn' by exposing the model to domain shift during training for learning more generalizable models. Zhao $et\ al.$ [48] proposed a Memory-based Multi-Source Meta-Learning (M^{3}L) framework, which overcomes the unstable meta-optimization by a memory-based and non-parametric identification loss.

The second category is ***Domain Alignment*** [14], which attempts to minimize the differences between source domains for pursuing the invariant features across domains. Jin $et~al.$ [14] propose a Style Normalization and Restitution (SNR) module to separate the identity-relevant and identity-irrelevant features by a dual causality loss constraint.

The third category is ***Mixture of Experts (MoE)*** [6]. Compared to other methods, MoE is superior in two aspects: (1) MoE learns diverse experts for different domains, which retains domain-specific knowledge for providing complementary characteristics. (2) MoE takes the target domain's inherent relevance w.r.t. diverse source domains into consideration for better generalization. Dai $et~al.$ [6] proposed a method called the relevance-aware mixture of experts (RaMoE), which adds a branch network (expert) for each source domain, and designs a voting network for integrating multiple experts. However, [6] suffers from a large model size with the increase of the number of source domains, which limits the application of the RaMoE. To tackle this problem, experts in our method share all the parameters except for the batch normalization layers.

Domain-Specific Batch Normalization. The statistics of BN vary in different domains. Therefore, mixing multiple source domains' statistics may be detrimental to improving generalizable performance [53]. To tackle this problem, domain-specific BN has been studied recently [21, 24, 29, 30]. Domain-specific BN works as constructing domain-specific classifiers but shares most of the parameters except for the BN layers. Different from prior methods, our method aggregates domain-specific predictions by a learnable aggregation network, which is able to measure the target domain's inherent relevance w.r.t. diverse source domains via normalization statistics.

3. Methodology

The structure of the META is illustrated in Fig. 2. After backbone, we design a global branch for extracting domain-invariant features and an expert branch for capturing domain-specific characteristics. Domain-specific BN layers are adopted in the backbone and the Exp-Block. In the expert branch, we develop a learnable aggregation network to adaptively integrate diverse domain experts via normalization statistics. Finally, we concatenate the features produced by the two branches for inference.

3.1. Preliminary

In almost all the prior DG ReID methods [1, 5, 32, 48], they share BN layers for all the source domains, which may neglect individual domains' discriminative characteristics and be detrimental to dealing with the domain gap [4, 6]. To leverage the complementary information of the source domains, inspired by [3, 4, 28], we adopt *domain-specific batch normalization* in META.

Let $X \in \mathbb{R}^{C \times H \times W}$ denotes a feature map, where C, H, W respectively indicate the number of channels, height and width. BN layer normalizes features by:

$$BN(X; \gamma, \beta, \mu, \sigma) = \gamma \cdot \frac{X - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta, \tag{1}$$

where $\mu \in \mathbb{R}^C$ and $\sigma \in \mathbb{R}^C$ are mean value and standard deviation respectively calculated with respect to a mini-batch, $\gamma \in \mathbb{R}^C$ and $\beta \in \mathbb{R}^C$ are affine parameters, and $\epsilon > 0$ is a small constant to avoid divided-by-zero. At training time, μ and σ are estimated by the moving average operation. Finally, the parameters μ, σ, γ, and β are able to represent the characteristics of the corresponding domain. We design individual BN layers for each source domain. Specifically, as shown in Fig. 2, backbone contains $K + 1$ BN layers. The BN-g is a global layer which is updated by the training data from all the source domains to help extract domain-invariant features. Other K BN layers are updated by the training data from the corresponding source domain to exploit domain-specific characteristics. The Exp-Block only contains K domain-specific BN layers.

Although we have exploited the complementary information of the source domains via *domain-specific batch normalization*, it is still challenging to approximate the population statistics of the unseen target domain because target domain data cannot be accessed at training time. To do this, we rely on IN statistics to capture the characteristics of the target domain. IN layers normalize features by:

$$IN(X; \gamma, \beta, \mu, \sigma) = \gamma \cdot \frac{X - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta. \tag{2}$$

Different from BN, here mean value μ and standard deviation σ are calculated with respect to each sample and each channel:

$$\mu = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} X, \tag{3}$$

$$\sigma = \sqrt{\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (X - \mu)^2 + \epsilon}. \tag{4}$$

At testing time, we calculate the IN statistics of the test samples to capture the characteristics of the unseen target domain.

3.2. Expert Branch

We expect those relevant source domains to contribute more valuable information than those less relevant domains. In this section, we explain how to measure the relevance of the target domain w.r.t. source domains via BN and IN statistics for integrating multiple experts. From Eq. (2) and

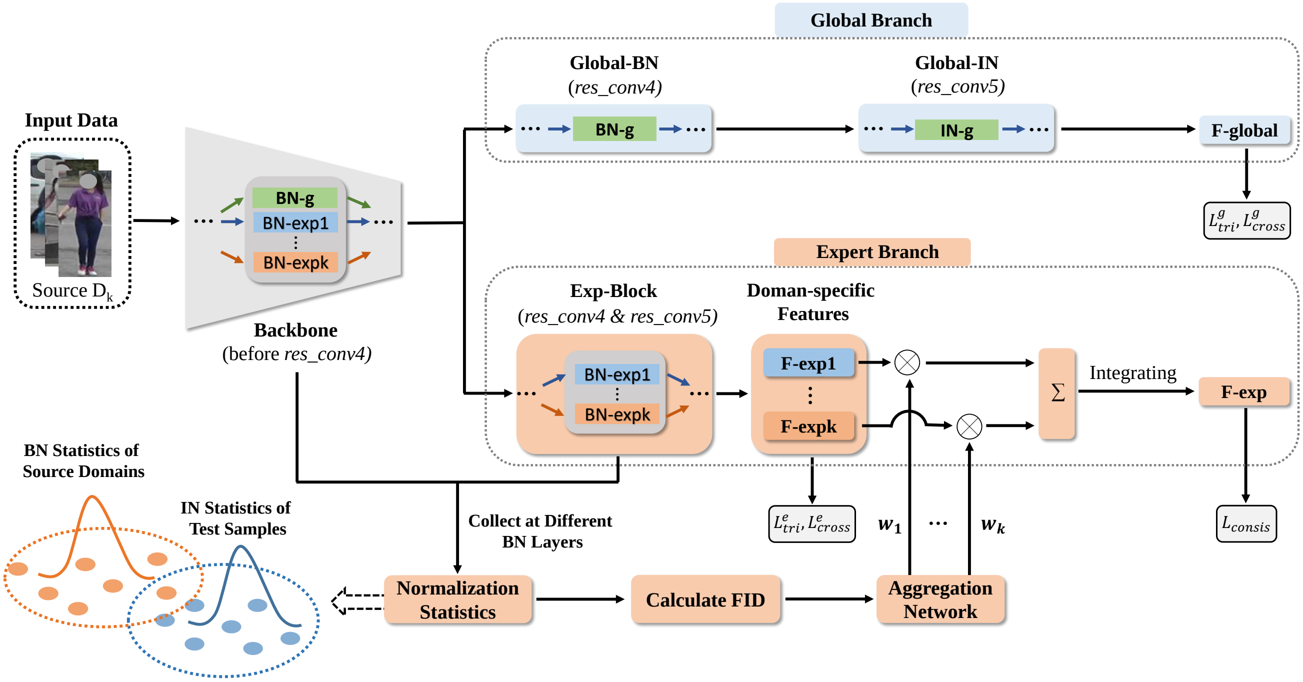


Figure 2. Overview of the META. '$\otimes$' is the operation of element-wise multiplication. '$\sum$' is a series of features' operation: element-wise division or summation. META is composed of a global branch for capturing domain-invariant features and an expert branch for exploiting complementary domain-specific information. The Exp-Block contains K domain-specific BN layers while the backbone contains K domain-specific BN layers and a global layer BN-g. BN-g is updated by the training data from all the source domains to help extract domain-invariant features. We replace the BN layers in the res_conv5 with IN layers to construct *Global-IN* for helping extract domain-invariant features. In the expert branch, we collect the IN statistics of the test samples and BN statistics of the source domains at different BN layers and calculate the *Fréchet Inception Distance* (FID) between them to measure the relevance of target domain w.r.t. source domains. Such relevance is leveraged by an aggregation network to adaptively integrate multiple experts. Finally, we concatenate F-global and F-exp for inference.

Eq. (1), we can see that IN is the degenerate case of BN with batch size equal to 1. META is built on such observation that BN and IN statistics are approximations of the same underlying distribution with different degrees of noise. Therefore, we can measure the relevance of the target domain w.r.t. source domains by comparing IN and BN statistics of them.

Specifically, we collect the BN statistics of source domains at different BN layers. We denote $D_k^{(l)} = (\mu_k^{(l)}, \sigma_k^{(l)^2})$ the BN statistics of source domain D_k at l-th BN layer. For each test sample x from an unseen target domain T_t, we forward propagate x through the network and calculate its IN statistics by Eq. (3) (4) at l-th BN layer as $T_t^{(l)} = (\mu_t^{(l)}, \sigma_t^{(l)^2})$. We select *Fréchet Inception Distance* (FID) to compute the distance between the target domain and source domains. Let $X \sim N(\mu_x, C_x)$ and $Y \sim N(\mu_y, C_y)$ be two normal distributions on R^n, with mean value $\mu_x, \mu_y \in R^n$ and covariance matrices $C_x, C_y \in R^{n\times n}$ respectively. The FID value is computed as:

$$\begin{aligned} FID(X,Y) &= \phi((\mu_x, C_x),(\mu_y, C_y)) \\ &= || \mu_x - \mu_y ||_2^2 + Tr(C_x + C_y - 2(C_x C_y)^{\frac{1}{2}}), \end{aligned} \tag{5}$$

where $Tr(\cdot)$ denotes the trace of the matrix and $|| \cdot ||$ denotes the Euclidean norm. Thereafter, we use Eq. (5) to calculate FID at different BN layers and concatenate them as:

$$\begin{aligned} R_k^t &= [r_k^{(1)}, r_k^{(2)}, ..., r_k^{(l)}] \in R^{1\times L}, \\ \text{where } r_k^{(l)} &= FID(D_k^{(l)}, T_t^{(l)}) \\ &= \phi((\mu_k^{(l)}, Diag(\sigma_k^{(l)^2})), (\mu_t^{(l)}, Diag(\sigma_t^{(l)^2}))), \end{aligned} \tag{6}$$

$r_k^{(l)}$ is the distance between the BN statistics of source domain D_k and IN statistics of target domain T_t at l-th layer. $Diag(\cdot)$ returns a square diagonal matrix with the elements of input vector on the main diagonal.

Then, we forward propagate R_k^t to an aggregation network, which is composed of two fully-connected layers, for computing the weight of domain-specific expert:

$$w_k = h_2(h_1(R_k^t)), \tag{7}$$

where h_1, h_2 are fully-connected layers of the aggregation network. During testing, we get the *F-exp* as a linear combination of the multiple experts:

$$F\text{-}exp(x) = \sum_{k=1}^{K} \frac{w_k f(x \mid k)}{\sum_j w_j}, \tag{8}$$

where $f(x \mid k)$ is the result of a forward pass of the k-th expert in the network. During training, we get the *F-exp* in another way, which will be introduced in Section 3.4. In this way, relevant source domains are able to contribute more valuable information than those less relevant domains for better generalization performance on the target domain.

3.3. Global Branch

We design a global branch to learn the domain-invariant features, which works as a complement to the domain-specific representations extracted by the expert branch for better generalizability. IN works on normalizing features with the statistics of individual instances, by which the domain-specific information could be filtered out from the content [7]. Inspired by this, we leverage IN layers in the global branch to capture the domain-invariant features.

Specifically, as shown in Fig. 2, the global branch is composed of the *Global-Bn* and *Global-In* blocks. *Global-Bn* block is the same as res_conv4. We replace all the BN layers in the res_conv5 with IN layers to build the *Global-In* block. Furthermore, training samples from all the source domains are used to update the global branch.

3.4. Training Policy

At training time, each training batch is composed of the training samples collected from the same source domain. Let x denotes the current training sample collected from source domain D_i ($1 \leq i \leq K$). We freeze all the BN layers except for the BN-g and i-th BN-exp. as shown in Fig. 2, we update the global branch by the triplet loss $\mathcal{L}_{tri}^{g}$ and cross-entropy loss $\mathcal{L}_{cross}^{g}$. Meanwhile, we optimize the i-th expert by the triplet loss $\mathcal{L}_{tri}^{e}$ and cross-entropy loss $\mathcal{L}_{cross}^{e}$. We combine the above losses as:

$$\mathcal{L}_{base} = \mathcal{L}_{tri}^{g} + \mathcal{L}_{cross}^{g} + \mathcal{L}_{tri}^{e} + \mathcal{L}_{cross}^{e}. \tag{9}$$

Algorithm 1: Training Procedure of META

```
Input: Training data $x$ from source domain $D_i$; MaxIters;
       MaxEpochs.
Output: Feature extractor $F_\theta(\cdot)$; Domain-specific experts
        $\{f(x \mid k)\}_{k=1}^{K}$.
Initialization;
for epoch=1 to MaxEpochs do
    for iter=1 to MaxIters do
        Domain-specific BN layers:
        Freeze all the BN layers except for the BN-g and
          $i$-th BN-exp;
        Global Branch:
        Update global branch by $\mathcal{L}_{tri}^{g}$ and $\mathcal{L}_{cross}^{g}$;
        Expert Branch:
        Update expert branch by $\mathcal{L}_{tri}^{e}$ and $\mathcal{L}_{cross}^{e}$;
        Aggregation Network:
        Combine $\{f(x \mid k)\}_{k=1,k\neq i}^{K}$ by Eq. (10) to
          produce F-exp;
        Update aggregation network by $\mathcal{L}_{consis}$ in
          Eq. (11);
    end
end
```

In addition, we adopt episodic training [16] which simulates the test process at training time to update the aggregation network. When x is input to the network, domain D_i is seemed as the 'unseen target domain' to the other $K-1$ domain-specific experts $\{f(x \mid k)\}_{k=1,k\neq i}^{K}$. We combine these $K-1$ domain experts to produce the representation *F-exp*, which is formulated as:

$$F\text{-}exp(x) = \sum_{k=1,k\neq i}^{K} \frac{w_k f(x \mid k)}{\sum_{j,j\neq i} w_j}, \; x \in D_i, \tag{10}$$

where w_k is the weight of k-th expert and $f(x \mid k)$ is the result of a forward pass of the k-th expert. To mimic embedding of D_i with *F-exp*, we propose a consistency loss to push the aggregated feature *F-exp* as discriminative as the feature $f(x \mid i)$ extracted by the i-th expert. The consistency loss is formulated as:

$$\mathcal{L}_{consis} = [\alpha_1 + \Gamma_{exp}^{+} - \Gamma_{i}^{+}]_+ + [\alpha_2 + \Gamma_{i}^{-} - \Gamma_{exp}^{-}]_+, \tag{11}$$

where α_1 and α_2 are margins, Γ_{exp}^{+} and Γ_{i}^{+} are hardest positive distances [11] of *F-exp* and $f(x \mid i)$ respectively, Γ_{exp}^{-} and Γ_{i}^{-} are hardest negative distances [11] of *F-exp* and $f(x \mid i)$ respectively. By minimizing Eq. (11), the aggregation network is learned to explicitly mimic the target domain via multiple experts. The total loss can be formulated as :

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_{consis}. \tag{12}$$

At test time, we combine K domain experts by Eq. (8) to produce *F-exp*, and concatenate it with *F-global* as the final representation. The overall training procedure is shown in Algorithm 1.

Table 1. The details of 9 datasets in the experiments.

Datasets	#IDs	#Images	#Cameras
Market1501 (M) [49]	1,501	32,217	6
MSMT17 (MS) [42]	4,101	126,441	15
CUHK02 (C2) [17]	1,816	7,264	10
CUHK03 (C3) [18]	1,467	14,096	2
CUHK-SYSU (CS) [43]	11,934	34,574	1
PRID [12]	749	949	2
GRID [23]	1,025	1,275	8
VIPeR [9]	632	1,264	2
iLIDs [50]	300	4,515	2

Table 2. Evaluation Protocols in the experiments. 'Full' denotes that both the training and testing sets are used for training.

	Training Sets	Testing Sets
Protocol-1	Full-(M+C2+C3+CS)	PRID,GRID, VIPeR,iLIDs
Protocol-2	M+MS+CS	C3
	M+CS+C3	MS
	MS+CS+C3	M
Protocol-3	Full-(M+MS+CS)	C3
	Full-(M+CS+C3)	MS
	Full-(MS+CS+C3)	M

4. Experiments

4.1. Datasets and Settings

Datsets. We conduct extensive experiments on 9 public ReID or person search datasets including Market1501 [49], MSMT17 [42], CUHK02 [17], CUHK03 [18], CUHK-SYSU [43], PRID [12], GRID [23], VIPeR [9], and iLIDs [50]. The details of these datasets are illustrated in Table 1. For CUHK03, we use the 'labeled' data as [6]. For simplicity, we denote MSMT17 as MS, Market1501 as M, CUHK02 as C2, CUHK03 as C3, and CUHK-SYSU as CS. We utilize Cumulative Matching Characteristics (CMC) and mean average precision (mAP) for evaluation.

Evaluation Protocols. Because DukeMTMC-reID [51], which was widely used in previous work [1, 5, 32, 48] on DG ReID, has been taken down, we set three new protocols for DG ReID, as shown in Table 2. For protocol-1, we use all the images in the source domains (*i.e.*, including training and testing sets) for training. For PRID, GRID, VIPeR, and iLIDS, following [6], the results are evaluated on the average of 10 repeated random splits of query and gallery sets. For protocol-2, we choose one domain from M+MS+CS+C3 for testing and the remaining three domains for training. As CS person search dataset only contains 1 camera, CS is not used for testing. The difference between protocol-2 and protocol-3 is that we use all the images in the source domains for training under protocol-3.

Implementation Details. We resize all the images to 256 × 128. ResNet50 [10] pretrained on ImageNet is used as our backbone. We set batch size to 64, including 16 identities and 4 images per identity. Similar to [6], we perform color jitter and discard random erasing for the data augmentation. We train the model for 120 epochs and adopt the warmup strategy in the first 500 iterations. The learning rate is initialized as $3e^{-4}$ and divided by 10 at the 40th and 70th epochs respectively. The margins α_1, α_2 in Eq. (11) are set to be zero.

4.2. Comparison with State-of-the-art Methods

Comparison under protocol-1. We compare our method with other state-of-the-arts under protocol-1, as shown in Table 3. We report some results of other methods which leverage DukeMTMC-reID [51] in the source domains, while we remove it from our training sets. Although we use fewer source domains, we still get the best performance. Specifically, from the results, we can find that META achieves the best performances on the PRID, GRID and VIPeR, while RaMoE [6] gives the highest points on the iLIDs dataset. META significantly outperforms other methods by at least 2.0% and 2.3% in average mAP and Rank-1 respectively.

Comparison under protocol-2 and protocol-3. We compare our method with other state-of-the-arts under protocol-2 and protocol-3, as shown in Table 4. 'Training Sets' denotes that only the training sets in the source domains are used for training and 'Full Images' denotes that all images in the source domains ($i.e.$ including training and testing sets) are leveraged at training time. The results show that META outperforms other methods by a large margin on all the datasets and under both protocols. Specifically, META surpasses other methods, on average, by at least 4.6% mAP, 5.6% Rank-1 and 6.3% mAP, 6.6% Rank-1 under protocol-2 and protocol-3 respectively. The results have shown our model's superiority in domain generalization.

4.3. Ablation Study

The effectiveness of the individual branches. We study ablation studies on the effectiveness of individual branches, as shown in the first, second, and last rows of Table 5. The experiment is conducted under protocol-3. We train our model without the global branch or expert branch for comparison. From the results, we can find that mAP drops 20.7%, 14.1% and 32.4% on the CUHK03, MSMT17 and Market1501 respectively when the global branch is discarded. The mAP also drops 13.5%, 3.9% and 4.6% on the CUHK03, MSMT17 and Market1501 respectively when the expert branch is discarded. The results have demonstrated the effectiveness of both the global and expert branches for improving domain generalization performance.

The effectiveness of aggregation network. We study ablation studies on the effectiveness of aggregation network, as shown in the third and last rows of Table 5. The

Table 3. Comparison with state-of-the-art methods under protocol-1. All the images in the source domains are used for training. 'M', 'D', 'MS', 'CS', 'C3' are the abbreviations of the Market1501, DukeMTMC-reID, MSMT17, CUHK-SYSU, and CUHK03 respectively. We report some results of other methods which leverage DukeMTMC-reID in the source domains, while we remove DukeMTMC-reID from our training sets. Although we use fewer source domains, we still get the best performance. '*' indicates that we re-implement this work based on the authors' code on Github. The best results are highlighted in bold.

Method	Reference	Source Domain	Target:PRID		Target:GRID		Target:VIPeR		Target:iLIDs		Average	
			mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
CrossGrad [31]	ICLR 2018	M+D +C2+C3+CS	28.2	18.8	16.0	8.96	30.4	20.9	61.3	49.7	34.0	24.6
Agg_PCB [35]	CVPR 2018		45.3	31.9	38.0	26.9	54.5	45.1	72.7	64.5	52.6	42.1
MLDG [15]	AAAI 2018		35.4	24.0	23.6	15.8	33.5	23.5	65.2	53.8	39.4	29.3
PPA [27]	TPAMI 2019		32.0	21.5	44.7	36.0	45.4	38.1	73.9	66.7	49.0	40.6
DIMN [33]	CVPR 2019		52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2	57.9	47.5
SNR [14]	CVPR 2020		66.5	52.1	47.7	40.2	61.3	52.9	89.9	84.1	66.4	57.3
RaMoE [6]	CVPR 2021		67.3	57.7	54.2	46.8	64.6	56.6	**90.2**	**85.0**	62.0	61.5
DMG-Net [2]	CVPR 2021		68.4	60.6	56.6	51.0	60.4	53.9	83.9	79.3	67.3	61.2
QAConv$_{50}$* [19]	ECCV 2020	M +C2+C3+CS	62.2	52.3	57.4	48.6	66.3	57.0	81.9	75.0	67.0	58.2
M^{3}L (ResNet-50)* [48]	CVPR 2021		65.3	55.0	50.5	40.0	68.2	60.8	74.3	65.0	64.6	55.2
MetaBIN* [5]	CVPR 2021		70.8	61.2	57.9	50.2	64.3	55.9	82.7	74.7	68.9	60.5
META (ours)		M +C2+C3+CS	**71.7**	**61.9**	**60.1**	**52.4**	**68.4**	**61.5**	83.5	79.2	**70.9**	**63.8**

Table 4. Comparison with state-of-the-art methods under protocol-2 and protocol-3. 'Training Sets' denotes that only the training sets in the source domains are used for training and 'Full Images' denotes that all images are leveraged at training time. 'M', 'MS', 'CS', 'C3' are the abbreviations of the Market1501, MSMT17, CUHK-SYSU, and CUHK03 respectively. '*' indicates that we re-implement this work based on the authors' code on Github. The best results are highlighted in bold.

Setting	Method	Reference	M+MS+CS→C3		M+CS+C3→MS		MS+CS+C3→M		Average	
			mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Protocol-2 (Training Sets)	SNR* [14]	CVPR2020	8.9	8.9	6.8	19.9	34.6	62.7	16.8	30.5
	QAConv$_{50}$* [19]	ECCV2020	25.4	24.8	16.4	45.3	63.1	83.7	35.0	51.3
	M^{3}L (ResNet-50)* [48]	CVPR2021	20.9	31.9	15.9	36.9	58.4	79.9	31.7	49.6
	M^{3}L (IBN-Net50)* [48]	CVPR2021	34.2	34.4	16.7	37.5	61.5	82.3	37.5	51.4
	MetaBIN* [5]	CVPR2021	28.8	28.1	17.8	40.2	57.9	80.1	34.8	49.5
	META (ours)		**36.3**	**35.1**	**22.5**	**49.9**	**67.5**	**86.1**	**42.1**	**57.0**
Protocol-3 (Full Images)	SNR* [14]	CVPR2020	17.5	17.1	7.7	22.0	52.4	77.8	25.9	39.0
	QAConv$_{50}$* [19]	ECCV2020	32.9	33.3	17.6	46.6	66.5	85.0	39.0	55.0
	M^{3}L (ResNet-50)* [48]	CVPR2021	32.3	33.8	16.2	36.9	61.2	81.2	36.6	50.6
	M^{3}L (IBN-Net50)* [48]	CVPR2021	35.7	36.5	17.4	38.6	62.4	82.7	38.5	52.6
	MetaBIN* [5]	CVPR2021	43.0	43.1	18.8	41.2	67.2	84.5	43.0	56.3
	META (ours)		**47.1**	**46.2**	**24.4**	**52.1**	**76.5**	**90.5**	**49.3**	**62.9**

Table 5. Ablation study on the effectiveness of individual components. The experiment is conducted under protocol-3. 'C3', 'MS', 'M' are the abbreviations of the CUHK03, MSMT17, and Market1501 respectively. The best results are highlighted in bold.

Method	Target: C3		Target: MS		Target: M		Average	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
w/o global branch	26.4	26.2	10.3	28.3	44.1	71.6	26.9	42.0
w/o expert branch	33.6	33.7	20.5	45.8	71.9	87.6	42.0	55.7
w/o aggregation network	46.0	45.5	23.3	50.9	75.1	89.6	48.1	62.0
META (ours)	**47.1**	**46.2**	**24.4**	**52.1**	**76.5**	**90.5**	**49.3**	**62.9**

experiment is conducted under protocol-3. *'w/o aggregation network'* denotes that we remove the aggregation network and directly integrate multiple experts with FID. The results show that the aggregation network gives the performance gains of 1.1%, 1.1% and 1.4% for mAP on CUHK03, MSMT17 and Market1501 respectively. The results have validated the effectiveness of the aggregation network for adaptively integrating diverse domain experts to mimic unseen target domain.

The impact of the Instance Normalization in the global branch. We evaluate the impact of IN for helping extract domain-invariant features in the global branch, as shown in Table 6. The experiment is conducted under

Table 6. Ablation study on the impact of the Instance Normalization in the global branch. The experiment is conducted under protocol-3. 'C3', 'MS', 'M' are the abbreviations of the CUHK03, MSMT17, and Market1501 respectively. The best results are highlighted in bold.

Method	Target: C3		Target: MS		Target: M	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Replace IN with BN	43.3	43.1	21.9	48.6	71.7	88.3
META (ours)	**47.1**	**46.2**	**24.4**	**52.1**	**76.5**	**90.5**

Table 7. Ablation study on the performance of the individual features. The experiment is conducted under protocol-3. 'C3', 'MS', 'M' are the abbreviations of the CUHK03, MSMT17, and Market1501 respectively. The best results are highlighted in bold.

Method	Target: C3		Target: MS		Target: M	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
F-global	46.9	46.0	24.1	52.0	76.4	90.3
F-exp	33.6	33.7	20.5	45.8	71.9	87.6
META (ours)	**47.1**	**46.2**	**24.4**	**52.1**	**76.5**	**90.5**

protocol-3. We replace IN in the *Global-IN* with BN for comparison. The results show that IN gives 3.8%, 2.5% and 4.8% higher points in mAP on CUHK03, MSMT17 and Market1501 respectively. The reason is that IN is able to filter out domain-specific information, which is beneficial to the domain-invariant representations extraction. The results have demonstrated the effectiveness of leveraging IN in the global branch to help extract domain-invariant features.

Performance of individual features. We study ablation studies on the performance of individual features, as shown in Table 7. The experiment is conducted under protocol-3. We separately inference with *F-global* and *F-exp* for comparison. The results show that *F-global* has a similar performance with META which concatenates *F-global* and *F-exp* for testing. We think the reason is that the expert branch is able to help the backbone extract more generalizable features, and therefore could improve the domain generalization performance of the global branch. As a result, it is feasible to only leverage the global branch during testing for faster inference.

The effectiveness of loss function components. We study ablation studies on the effectiveness of loss function components, as shown in Table 8. The experiment is conducted under protocol-2. $\mathcal{L}_{cross}$ and $\mathcal{L}_{tri}$ indicate that we replace $\mathcal{L}_{consis}$ with cross-entropy loss and triplet loss respectively to update the aggregation network. From the first and fourth rows, we can find that $\mathcal{L}_{consis}$ gives performance gains of 2.2% and 4.3% for mAP and Rank-1 accuracy respectively. From the last three rows, we can find that $\mathcal{L}_{consis}$ achieves the best performance, which surpasses $\mathcal{L}_{cross}$ and $\mathcal{L}_{tri}$ by 8.4% and 1.7% Rank-1 accuracy respectively. The results have demonstrated the effectiveness of our proposed $\mathcal{L}_{consis}$.

4.4. Discussion & Limitation

Although META achieves satisfying results on DG ReID, there are still some limitations: (1) META leverages domain-specific BN layers for exploiting diversified characteristics of source domains. However, some backbones lack BN layers, such as Transformer [38] and Inception Network [37]. How to combine META with these backbones needs to be studied further. (2) All the experts are input to the aggregation network during integration. However, this may be time-consuming when the number of source domains increases. How to choose which experts to be aggregated could be studied further.

Table 8. Ablation study on the loss function components. The experiment is conducted under protocol-2. $\mathcal{L}_{cross}$ and $\mathcal{L}_{tri}$ indicate that we replace $\mathcal{L}_{consis}$ with cross-entropy loss and triplet loss respectively to update the aggregation network. The best results are highlighted in bold.

$\mathcal{L}_{base}$	$\mathcal{L}_{cross}$	$\mathcal{L}_{tri}$	$\mathcal{L}_{consis}$	Target: MSMT17	
				mAP	Rank-1
√				20.3	45.6
√	√			17.8	41.5
√		√		21.4	48.2
√			√	**22.5**	**49.9**

5. Conclusion

This paper presents a new approach called Mimicking Embedding via oThers' Aggregation (META) for Domain generalizable (DG) person re-identification (ReID). META is a lightweight ensemble of multiple experts sharing all the parameters except for the domain-specific BN layers. Besides multiple experts, META leverages Instance Normalization (IN) and introduces it into a global branch to pursue invariant features across domains. Meanwhile, META develops an aggregation network to adaptively integrate multiple experts with the relevance of an unseen target sample w.r.t. source domains via normalization statistics. Benefiting from a proposed consistency loss and an episodic training algorithm, we can expect META to mimic embedding for a truly unseen target domain. Extensive experiments demonstrate that META surpasses state-of-the-art DG ReID methods by a large margin.

6. Ethical concerns

As for positive impact, we focus on the DG ReID problem. We need not collect data from the target domain and only leverage existing data to train the model. Hence, our work has the potential to mitigate ethical concerns about collecting vast volumes of pedestrian data. However, exploiting ReID system may violate people's privacy. Because ReID systems typically (but not always) need to collect data from unauthorized surveillance, which means not all the pedestrians were aware they were being recorded. As a result, governments and officials must go to great lengths to create stringent regulations and legislation governing the use of ReID technology.

References

[1] Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, and Ling-Yu Duan. Person30k: A dual-meta generalization network for person re-identification. In *CVPR*, 2021. 1, 2, 3, 6

[2] Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, and Ling-Yu Duan. Person30k: A dual-meta generalization network for person re-identification. In *CVPR*, 2021. 7

[3] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding. Unsupervised multi-source domain adaptation for person re-identification. In *CVPR*, 2021. 3

[4] W. G. Chang, T. You, S. Seo, S. Kwak, and B. Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, 2019. 3

[5] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *CVPR*, 2021. 1, 2, 3, 6, 7

[6] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, 2021. 1, 2, 3, 6, 7

[7] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv*, 2016. 2, 5

[8] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, 2019. 1, 2

[9] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 6

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

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 5

[12] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102, 2011. 6

[13] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2

[14] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, 2020. 1, 2, 3, 7

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 7

[16] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. In *ICCV*, 2019. 2, 5

[17] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 6

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 6

[19] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *ECCV*, 2020. 7

[20] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, 2018. 2

[21] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE transactions on medical imaging*, 39(9):2713–2724, 2020. 3

[22] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, 2021. 1

[23] 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. 6

[24] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Robust place categorization with deep domain generalization. *IEEE Robotics and Automation Letters*, 3(3):2093–2100, 2018. 3

[25] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, pages 542–551, 2019. 1

[26] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *CVPR*, 2018. 2

[27] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018. 7

[28] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *arXiv*, 2020. 2, 3

[29] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *arXiv*, 2020. 3

[30] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, 2020. 3

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

[32] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, 2019. 1, 2, 3, 6

[33] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, 2019. 7

[34] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017. 2

[35] Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *PAMI*, 2019. 7

[36] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 2

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 8

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 8

[39] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018. 1

[40] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018. 2

[41] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 2

[42] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 6

[43] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv*, 2016. 6

[44] Boqiang Xu, Lingxiao He, Xingyu Liao, Wu Liu, Zhenan Sun, and Tao Mei. Black re-id: A head-shoulder descriptor for the challenging problem of person re-identification. In *ACM MM*, 2020. 1

[45] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, 2019. 2

[46] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*, 2020. 1, 2

[47] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 2

[48] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, 2021. 1, 2, 3, 6, 7

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

[50] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, volume 2, pages 1–11, 2009. 6

[51] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 6

[52] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 1, 2

[53] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv*, 2021. 3