Mimic Embedding via Adaptive Aggregation: Learning Generalizable Person Re-identification

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

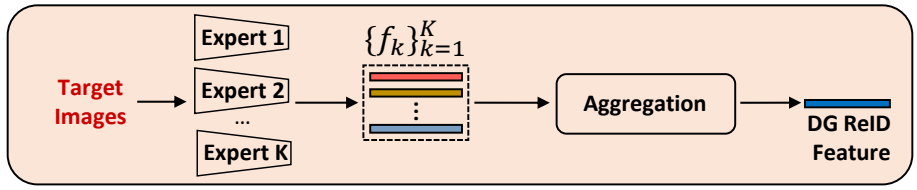
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[3] Longfor Inc.
boqiang.xu@cripac.ia.ac.cn
{liangjian92, xiaomingzhidao1}@gmail.com
znsun@nlpr.ia.ac.cn
*Corresponding author

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 Mimic Embedding via adapTive Aggregation (META) for DG person ReID. To avoid the large model size, experts in META do not adopt 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 module to adaptively integrate multiple experts for mimicking unseen target domain. Benefiting from a proposed consistency loss and an episodic training algorithm, META is expected to mimic embedding for a truly unseen target domain. Extensive experiments verify that META surpasses state-of-the-art DG person ReID methods by a large margin. Our code is available at https://github.com/xbq1994/META.

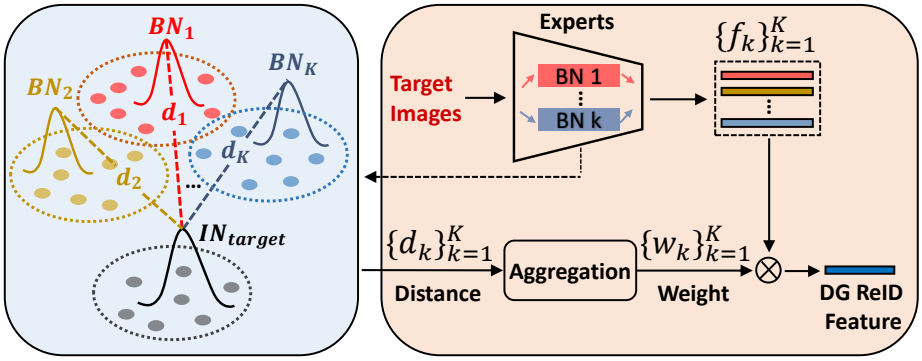
Keywords: Domain generalization; Person re-identification

1 Introduction

Person re-identification (ReID) aims at retrieving persons of the same identity across non-overlapping cameras. Many prior works [39,29,43,25,42,10] have been



(a) Prior MoE-based DG ReID Method



(b) Our Method (META)

Fig. 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 samples w.r.t. source domains. Such distances $\{d_k\}_{k=1}^K$ are exploited by an aggregation module to adaptively integrate multiple experts.

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 [40]. To avoid this, recent efforts are devoted to domain adaptive (DA) ReID [49,44,7] and domain generalizable (DG) ReID [45,4,5,17]. 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 person ReID problem.

Most of the prior DG ReID methods [45,4,35,1,17] assume an identical model for different domains. However, 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) [16] 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 Mimic Embedding via adapTive Aggregation (METa), as shown in Fig. 1b. Batch Normalization (BN) statistics are computed on-the-fly during training and can be seen as statistics of the characteristics of individual domain [32]. 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) [6], 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 target samples w.r.t. source domains. Taking the relevance as input, we further devise a small aggregation module 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 [19] which simulates the test process at training time for updating the aggregation module. 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 module 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 METa, a novel 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 module, 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 [38,39,36,23,11] have been proposed to improve the ReID performance. 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 [40]. To tackle this problem, some researchers start to study the unsupervised domain adaption (UDA) methods [49,44,7]. 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 [45,4,5,17] 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 [45,4,35,1]. 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*. [45] proposed a Memory-based Multi-Source Meta-Learning (M^3L) framework, which overcomes the unstable meta-optimization by a memory-based and non-parametric identification loss.

The second category is Domain Alignment [17], which attempts to minimize the differences between source domains for pursuing the invariant features across domains. Jin *et al*. [17] 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) [5]. MoE learns diverse experts for different domains and takes the target domain's inherent relevance w.r.t. diverse source domains into consideration for better generalization. Dai *et al*. [5] 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, [5] 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

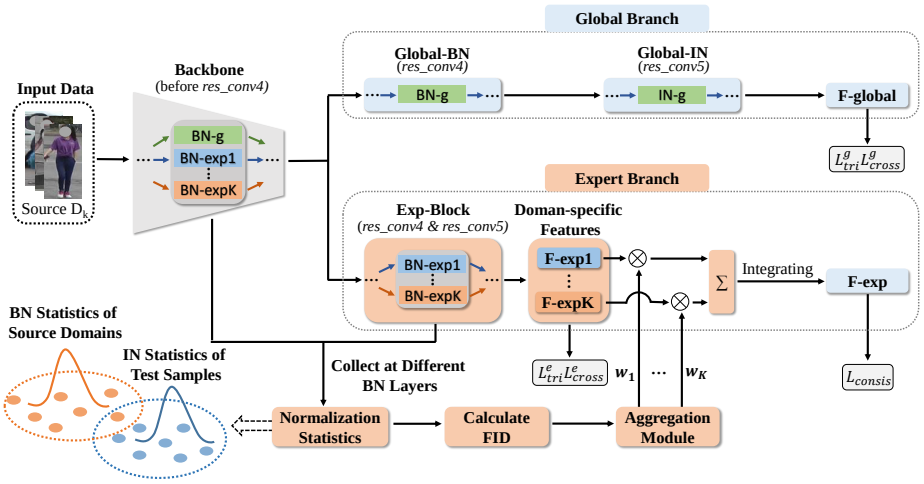


Fig. 2: Overview of the proposed META. '⊗' is the operation of element-wise multiplication. '$\sum$' is a series of features' operation: element-wise division or summation. META is composed of a global branch for capturing domain-invariant features and an expert branch for exploiting complementary domain-specific 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. We replace the BN layers in the *res_conv*5 with IN layers to construct Global-IN. K domain-specific BN layers are updated by their corresponding source domain's data to capture domain-specific characteristics while BN-g and IN-g are updated by the training data from all the source domains to help 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 samples w.r.t. source domains. Such relevance is leveraged by an aggregation module to adaptively integrate multiple experts. Finally, we concatenate F-global and F-exp for inference.

detrimental to improving generalizable performance [50]. To tackle this problem, domain-specific BN has been studied recently [33,24,32,28]. Domain-specific BN works as constructing domain-specific classifiers but shares most of the parameters except for the BN layers.

3 Methodology

Typically, we are provided with K source domains $\{D_k\}_{k=1}^{K}$ for training a DG ReID model, which have completely disjoint label spaces. In the testing phase, we directly test on unseen target domains without additional model updating. The structure of the META is illustrated in Fig. 2.

3.1 Preliminary

In almost all the prior DG ReID methods [45,4,35,1], 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 [5,3]. To leverage the complementary information of the source domains, inspired by [32,3,2], we adopt *domain-specific batch normalization* in METAL.

Let $X_k \in \mathbb{R}^{N\times C\times H\times W}$ denotes a feature map extracted from source domain D_k, where N, C, H, W respectively indicate the batch size, the number of channels, the height, and the width. BN layer normalizes features by:

$$BN(X_k) = \gamma_k^{bn} \cdot \frac{X_k - \mu_k^{bn}}{\sqrt{{\sigma_k^{bn}}^2 + \epsilon}} + \beta_k^{bn}, \tag{1}$$

where $\gamma_k^{bn} \in \mathbb{R}^C$ and $\beta_k^{bn} \in \mathbb{R}^C$ are affine parameters, $\epsilon > 0$ is a small constant to avoid divided-by-zero. $\mu_k^{bn} \in \mathbb{R}^C$ and $\sigma_k^{bn} \in \mathbb{R}^C$ are respectively mean value and standard deviation calculated with respect to a mini-batch and each channel:

$$\mu_k^{bn} = \frac{\sum_n \sum_{h,w} X_k}{N \cdot H \cdot W} \quad \text{and} \quad \sigma_k^{bn} = \sqrt{\frac{\sum_n \sum_{h,w} (X_k - \mu_k^{bn})^2}{N \cdot H \cdot W}}. \tag{2}$$

μ_k^{bn} and σ_k^{bn} are updated by the moving average operation [15] at training time and fixed during inference. We design individual BN layers for each source domain. Specifically, as shown in Fig. 2, Exp-Block contains K domain-specific BN layers, which are updated by the training data from the corresponding source domain to exploit domain-specific characteristics. Besides K domain-specific BN layers, another global layer BN-g is introduced in the backbone and global branch, which is updated by the training data from all the source domains to help extract domain-invariant features.

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, at testing time, we rely on IN statistics to capture the characteristics of the target samples. Given an example from target domain T_t, IN layers normalize features by:

$$IN(X_t) = \gamma_t^{in} \cdot \frac{X_t - \mu_t^{in}}{\sqrt{{\sigma_t^{in}}^2 + \epsilon}} + \beta_t^{in}. \tag{3}$$

Different from BN, mean value μ_t^{in} and standard deviation σ_t^{in} here are calculated with respect to each sample and each channel:

$$\mu_t^{in} = \frac{\sum_{h,w} X_t}{H \cdot W} \quad \text{and} \quad \sigma_t^{in} = \sqrt{\frac{\sum_{h,w} (X_t - \mu_t^{in})^2}{H \cdot W}}. \tag{4}$$

In the next section, we explain how to measure the relevance of the target samples w.r.t. source domains via BN and IN statistics.

3.2 Expert Branch in META

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 samples w.r.t. source domains via BN and IN statistics for integrating multiple experts. From Eq. (1)-Eq. (4), we can see that IN is the degenerate case of BN with batch size N equal to 1. META is built on such observation that BN and IN statistics are both approximations of Gaussian distributions (*i.e.*, they are comparable) and have potential to reflect the properties of the source domains and target samples respectively. Therefore, we can measure the relevance of the target samples 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. Considering a source domain D_k, we denote $D_k^{(l)} = (\mu_k^{bn(l)}, {\sigma_k^{bn(l)}}^2)$ the BN statistics at l-th layer of k-th BN-exp. For each test sample x_t from an unseen target domain T_t, we forward propagate x_t through the network and calculate its IN statistics by Eq. (4) at l-th layer of k-th BN-exp as $T_t^{(l)} = (\mu_t^{in(l)}, {\sigma_t^{in(l)}}^2)$. We adopt *Fréchet Inception Distance* (FID) [13] to compute the distance between the BN and IN statistics at l-th layer as:

$$\begin{aligned} r_{k,t}^{(l)} &= FID((\mu_k^{bn(l)}, {\sigma_k^{bn(l)}}^2), (\mu_t^{in(l)}, {\sigma_t^{in(l)}}^2)) \\ &= \|\mu_k^{bn(l)} - \mu_t^{in(l)}\|_2^2 + Tr(C_k^{(l)} + C_t^{(l)} - 2(C_k^{(l)} C_t^{(l)})^{\frac{1}{2}}), \\ \text{where } C_k^{(l)} &= Diag({\sigma_k^{bn(l)}}^2),\ C_t^{(l)} = Diag({\sigma_t^{in(l)}}^2), \end{aligned} \tag{5}$$

and $Diag(\cdot)$ returns a square diagonal matrix with the elements of input vector on the main diagonal. $r_{k,t}^{(l)}$ denotes the distance between the BN statistics of source domain D_k and IN statistics of test sample from target domain T_t at l-th layer, $||\cdot||$ denotes the Euclidean norm, and $Tr(\cdot)$ denotes the trace of the matrix. Thereafter, we concatenate $r_{k,t}^{(l)}$ at every layer as:

$$R_k^t = [r_{k,t}^{(1)}, r_{k,t}^{(2)}, ..., r_{k,t}^{(L)}] \in \mathbb{R}^{1\times L}. \tag{6}$$

Then, we forward propagate R_k^t to an aggregation module $h : \mathbb{R}^L \to \mathbb{R}$ for computing the weight of domain-specific expert:

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

where h consists of two fully-connected layers. The aggregation module further enhances the domains' relevance measure by adopting a learnable module. During testing, we get the *F-exp* as a linear combination of the multiple experts:

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

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 in META

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 [6]. Inspired by this, we leverage IN layers in the global branch to capture the domain-invariant features.

The global branch is designed based on the findings from [30] that adding IN layers after BN layers could significantly improve the domain generalization performance of the model. 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 $(1 \leq i \leq K)$. As shown in Fig. 2, we freeze all the BN layers except for the BN-g and i-th BN-exp. We update the global branch by the triplet loss [12] $\mathcal{L}^g_{tri}$ and cross-entropy loss $\mathcal{L}^g_{cross}$. Meanwhile, we optimize the i-th expert by the triplet loss [12] $\mathcal{L}^e_{tri}$ and cross-entropy loss $\mathcal{L}^e_{cross}$. Combining these losses above together, we have the following overall objective:

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

In addition, we adopt episodic training [19] which simulates the test process at training time to update the aggregation module. 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 $\{f(x \mid k)\}^K_{k=1,k\neq i}$. We combine these $K-1$ domain experts to produce the representation *F-exp*, which is formulated as:

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

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:

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

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 L_tri^g and L_cross^g;
        Expert Branch:
        Update expert branch by L_tri^e and L_cross^e;
        Aggregation Module:
        Combine {f(x | k)}_{k=1,k≠i}^K by Eq. (10) to produce F-exp;
        Update aggregation module by L_consis in Eq. (11);
    end
end
```

where α_1 and α_2 are margins, Γ_{exp}^{+} and Γ_{i}^{+} are hardest positive distances [12] of *F-exp* and $f(x \mid i)$ respectively, Γ_{exp}^{-} and Γ_{i}^{-} are hardest negative distances [12] of *F-exp* and $f(x \mid i)$ respectively, $[z]_+$ equals to $max(z, 0)$. By minimizing Eq. (11), the aggregation module 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.

4 Experiments

4.1 Datasets and Settings

Datsets. We conduct extensive experiments on 9 public ReID or person search datasets including Market1501 [46], MSMT17 [40], CUHK02 [20], CUHK03 [21], CUHK-SYSU [41], PRID [14], GRID [26], VIPeR [8], and iLIDs [47]. The details of these datasets are illustrated in Table 1. For CUHK03, we use the 'labeled' data as [5]. 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 [48], which was widely used in previous work [45,4,35,1] 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

Table 1: Summary of all the datasets.

Datasets	#IDs	#Images	#Cameras
Market1501 (M) [46]	1,501	32,217	6
MSMT17 (MS) [40]	4,101	126,441	15
CUHK02 (C2) [20]	1,816	7,264	10
CUHK03 (C3) [21]	1,467	14,096	2
CUHK-SYSU (CS) [41]	11,934	34,574	1
PRID [14]	749	949	2
GRID [26]	1,025	1,275	8
VIPeR [8]	632	1,264	2
iLIDs [47]	300	4,515	2

Table 2: Evaluation protocols.

	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

Table 3: Comparison with state-of-the-art methods under protocol-1. All the images in the source domains are used for training. The illustration of abbreviations is shown in Table 1. 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 (in **bold red**), the second best (in *italic blue*).

Method	Source Domains	→PRID		→GRID		→VIPeR		→iLIDs		Average	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
CrossGrad [34]	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 [37]		45.3	31.9	38.0	26.9	54.5	45.1	72.7	64.5	52.6	42.1
MLDG [18]		35.4	24.0	23.6	15.8	33.5	23.5	65.2	53.8	39.4	29.3
PPA [31]		32.0	21.5	44.7	36.0	45.4	38.1	73.9	66.7	49.0	40.6
DIMN [35]		52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2	57.9	47.5
SNR [17]		66.5	52.1	47.7	40.2	61.3	52.9	*89.9*	*84.1*	66.4	57.3
RaMoE [5]		67.3	57.7	54.2	46.8	64.6	56.6	**90.2**	**85.0**	62.0	*61.5*
DMG-Net [1]		68.4	60.6	56.6	*51.0*	60.4	53.9	83.9	79.3	67.3	61.2
$QAConv_{50}$ [22]*	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^3L(ResNet-50) [45]*		65.3	55.0	50.5	40.0	*68.2*	*60.8*	74.3	65.0	64.6	55.2
MetaBIN [4]*		*70.8*	*61.2*	*57.9*	50.2	64.3	55.9	82.7	74.7	*68.9*	60.5
META		**71.7**	**61.9**	**60.1**	**52.4**	**68.4**	**61.5**	83.5	79.2	**70.9**	**63.8**

all the images in the source domains (*i.e.*, including training and testing sets) for training. For PRID, GRID, VIPeR, and iLIDS, following [5], 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 the 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 [9] pretrained on ImageNet is used as our backbone. We set batch size to 64, including 16 identities and 4 images per identity. Similar to [5], 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

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. The illustration of abbreviations is shown in Table 1. '*' indicates that we re-implement this work based on the authors' code on Github. The best (in **bold red**), the second best (in *italic blue*).

Method	Setting	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
SNR* [17]	Protocol-2 (Training Sets)	8.9	8.9	6.8	19.9	34.6	62.7	16.8	30.5
$QAConv_{50}$ [22]*		25.4	24.8	16.4	*45.3*	*63.1*	*83.7*	35.0	51.3
M^3L (ResNet-50) [45]*		20.9	31.9	15.9	36.9	58.4	79.9	31.7	49.6
M^3L (IBN-Net50) [45]*		*34.2*	*34.4*	16.7	37.5	61.5	82.3	*37.5*	*51.4*
MetaBIN [4]*		28.8	28.1	*17.8*	40.2	57.9	80.1	34.8	49.5
META		**36.3**	**35.1**	**22.5**	**49.9**	**67.5**	**86.1**	**42.1**	**57.0**
SNR* [17]	Protocol-3 (Full Images)	17.5	17.1	7.7	22.0	52.4	77.8	25.9	39.0
$QAConv_{50}$* [22]		32.9	33.3	17.6	*46.6*	66.5	*85.0*	39.0	55.0
M^3L (ResNet-50) [45]*		32.3	33.8	16.2	36.9	61.2	81.2	36.6	50.6
M^3L (IBN-Net50) [45]*		35.7	36.5	17.4	38.6	62.4	82.7	38.5	52.6
MetaBIN [4]*		*43.0*	*43.1*	*18.8*	41.2	*67.2*	84.5	*43.0*	*56.3*
META		**47.1**	**46.2**	**24.4**	**52.1**	**76.5**	**90.5**	**49.3**	**62.9**

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 0.1.

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 [48] 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 [5] 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

Table 5: Ablation study on the effectiveness of individual components and the design of 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		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 module	46.0	45.5	23.3	50.9	75.1	89.6	48.1	62.0
BN-BN	43.3	43.1	21.9	48.6	71.7	88.3	45.6	60.0
BN-IBN [30]	45.2	44.0	22.7	50.2	73.2	89.8	47.0	61.3
IN-IN	41.5	40.2	18.7	46.0	68.3	86.7	42.8	57.6
META	**47.1**	**46.2**	**24.4**	**52.1**	**76.5**	**90.5**	**49.3**	**62.9**

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. Furthermore, we visualize the features extracted by different branches via t-SNE [27], as shown in Fig. 3(a). Different colors denote various IDs. We find that the expert branch pushes features from different IDs away while the global branch pulls the features from same ID closer. Thus, both branches are integrated for better ReID performance.

The effectiveness of aggregation module. We study ablation studies on the effectiveness of aggregation module, as shown in the third and last rows of Table 5. The experiment is conducted under protocol-3. *'w/o aggregation module'* denotes that we remove the aggregation module and directly integrate multiple experts with FID. The results show that the aggregation module 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 module for adaptively integrating diverse domain experts to mimic unseen target domain.

Table 6: Ablation study on the performance of the individual features under protocol-3.

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	42.9	42.0	10.2	28.7	45.7	72.3
META	**47.1**	**46.2**	**24.4**	**52.1**	**76.5**	**90.5**

Table 7: Ablation study on loss functions under protocol-3.

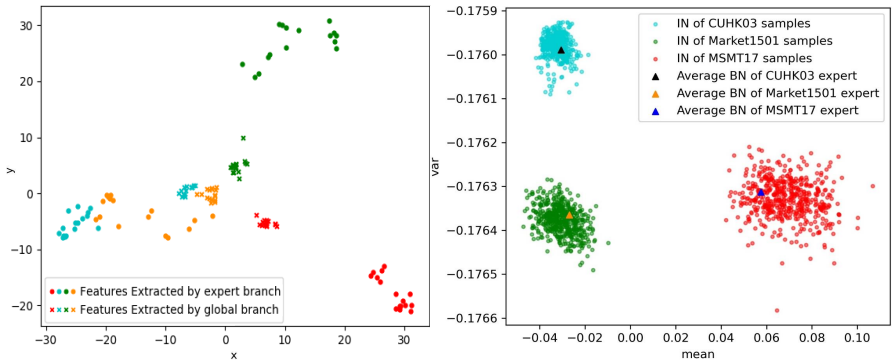
$\mathcal{L}_{base}$	$\mathcal{L}_{cross}$	$\mathcal{L}_{tri}$	$\mathcal{L}_{consis}$	Target: MSMT17	
				mAP	Rank-1
√				21.2	48.4
√	√			23.5	50.9
√		√		22.8	50.4
√			√	**24.4**	**52.1**

The design of global branch. The global branch is designed based on the findings from [30] that adding IN layers after BN layers could significantly improve the domain generalization performance of the model. We compare our design with other architectures of the global branch, as shown in the last four rows of Table 5. The experiment is conducted under protocol-3. We respectively replace IN in the *Global-IN* with BN and IBN [30], and replace BN in the *Global-BN* with IN for comparison. The results show that our design achieve the best results, surpassing other architectures by 2.9%, 1.6% and 5.3% respectively in average Rank-1. The results have demonstrated the effectiveness of our design of 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 6. 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 7. The experiment is conducted under protocol-3. $\mathcal{L}_{base}$ is defined in Eq. (9) for training the global and expert branch. $\mathcal{L}_{cross}$ and $\mathcal{L}_{tri}$ indicate that we replace $\mathcal{L}_{consis}$ with cross-entropy loss and triplet loss respectively to update the aggregation module. From the first and fourth rows, we can find that $\mathcal{L}_{consis}$ gives performance gains of 3.2% and 3.7% 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 1.2% and 1.7% Rank-1 accuracy respectively. The results have demonstrated the effectiveness of our proposed $\mathcal{L}_{consis}$.

The justification for calculating FID between BN and IN statistics. Both BN and IN can be seen as approximations of different Gaussian distributions, thus we can simply adopt FID to measure the difference between them. We expect through our learning scheme, BN and IN statistics could reflect the properties of the source and target domain respectively. Fig. 3(b) plots the average BN of multiple experts and IN statistics of samples from different domains via t-SNE [27]. The horizontal and vertical axes represent the mean and



(a) Visualization of different branches (b) Visualization of statistics

Fig. 3: (a) Visualization of the features extracted by different branches. Various colors denote different IDs. (b) Visualization of statistics, the results illustrate the justification for calculating FID between BN and IN statistics.

standard deviation of the statistics respectively. The result shows that different domain clusters can be divided by their IN statistics. Additionally, IN statistics of the samples are closer to the average BN of the expert from the same domain. The result illustrates the justification for calculating FID between BN and IN statistics.

5 Conclusion

This paper presents a new approach called Mimic Embedding via adapTive 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 module to adaptively integrate multiple experts with the relevance of an unseen target sample w.r.t. source domains via normalization statistics. Extensive experiments demonstrate that META surpasses state-of-the-art DG ReID methods by a large margin.

6 Acknowledgement

The authors would like to thank reviewers for providing valuable suggestions to improve this paper. This work is supported by the National Natural Science Foundation of China (Grant No. U1836217) and the Beijing Nova Program under Grant Z211100002121108.

References

1. Bai, Y., Jiao, J., Ce, W., Liu, J., Lou, Y., Feng, X., Duan, L.Y.: Person30k: A dual-meta generalization network for person re-identification. In: CVPR (2021) 2, 4, 6, 9, 10
2. Bai, Z., Wang, Z., Wang, J., Hu, D., Ding, E.: Unsupervised multi-source domain adaptation for person re-identification. In: CVPR (2021) 6
3. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: CVPR (2019) 6
4. Choi, S., Kim, T., Jeong, M., Park, H., Kim, C.: Meta batch-instance normalization for generalizable person re-identification. In: CVPR (2021) 2, 4, 6, 9, 10, 11
5. Dai, Y., Li, X., Liu, J., Tong, Z., Duan, L.Y.: Generalizable person re-identification with relevance-aware mixture of experts. In: CVPR (2021) 2, 4, 6, 9, 10, 11
6. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv (2016) 3, 8
7. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: ICCV (2019) 2, 4
8. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV (2008) 9, 10
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 10
10. He, L., Liang, J., Li, H., Sun, Z.: Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In: CVPR (2018) 1
11. He, L., Liu, W., Liang, J., Zheng, K., Liao, X., Cheng, P., Mei, T.: Semi-supervised domain generalizable person re-identification. arXiv (2021) 4
12. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv (2017) 8, 9
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017) 7
14. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian Conference on Image Analysis (2011) 9, 10
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015) 6
16. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3**(1), 79–87 (1991) 3
17. Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification. In: CVPR (2020) 2, 4, 10, 11
18. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Learning to generalize: Meta-learning for domain generalization. In: AAAI (2018) 10
19. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., Hospedales, T.M.: Episodic training for domain generalization. In: ICCV (2019) 3, 8
20. Li, W., Wang, X.: Locally aligned feature transforms across views. In: CVPR (2013) 9, 10
21. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014) 9, 10
22. Liao, S., Shao, L.: Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In: ECCV (2020) 10, 11

23. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: CVPR (2018) 4
24. Liu, Q., Dou, Q., Yu, L., Heng, P.A.: Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE Transactions on Medical Imaging **39**(9), 2713–2724 (2020) 5
25. Liu, X., Zhang, P., Yu, C., Lu, H., Yang, X.: Watching you: Global-guided reciprocal learning for video-based person re-identification. In: CVPR (2021) 1
26. Loy, C.C., Xiang, T., Gong, S.: Time-delayed correlation analysis for multi-camera activity understanding. International Journal of Computer Vision **90**(1), 106–129 (2010) 9, 10
27. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008) 12, 13
28. Mancini, M., Bulo, S.R., Caputo, B., Ricci, E.: Robust place categorization with deep domain generalization. IEEE Robotics and Automation Letters **3**(3), 2093–2100 (2018) 5
29. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: ICCV (2019) 1
30. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV (2018) 8, 12, 13
31. Qiao, S., Liu, C., Shen, W., Yuille, A.L.: Few-shot image recognition by predicting parameters from activations. In: CVPR (2018) 10
32. Segu, M., Tonioni, A., Tombari, F.: Batch normalization embeddings for deep domain generalization. arXiv (2020) 3, 5, 6
33. Seo, S., Suh, Y., Kim, D., Kim, G., Han, J., Han, B.: Learning to optimize domain specific normalization for domain generalization. In: ECCV (2020) 5
34. Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S.: Generalizing across domains via cross-gradient training. arXiv (2018) 10
35. Song, J., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: CVPR (2019) 2, 4, 6, 9, 10
36. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV (2017) 4
37. Sun, Y., Zheng, L., Li, Y., Yang, Y., Tian, Q., Wang, S.: Learning part-based convolutional features for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(3), 902–917 (2019) 10
38. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV (2018) 4
39. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM MM (2018) 1, 4
40. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR (2018) 2, 4, 9, 10
41. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. arXiv (2016) 9, 10
42. Xu, B., He, L., Liang, J., Sun, Z.: Learning feature recovery transformer for occluded person re-identification. IEEE Transactions on Image Processing **31**, 4651–4662 (2022) 1
43. Xu, B., He, L., Liao, X., Liu, W., Sun, Z., Mei, T.: Black re-id: A head-shoulder descriptor for the challenging problem of person re-identification. In: ACM MM (2020) 1

44. Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., Tian, Y.: Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: CVPR (2020) 2, 4
45. Zhao, Y., Zhong, Z., Yang, F., Luo, Z., Lin, Y., Li, S., Sebe, N.: Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In: CVPR (2021) 2, 4, 6, 9, 10, 11
46. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015) 9, 10
47. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: BMVC. pp. 1–11 (2009) 9, 10
48. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV (2017) 9, 11
49. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: CVPR (2019) 2, 4
50. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. arXiv (2021) 5