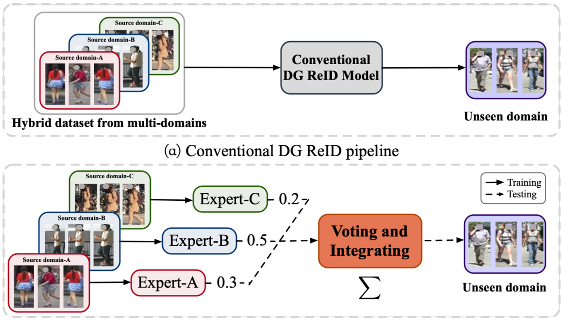
Generalizable Person Re-identification with Relevance-aware Mixture of Experts

Yongxing Dai[1] Xiaotong Li[1] Jun Liu[2] Zekun Tong[3] Ling-Yu Duan[1,4*]
[1] National Engineering Lab for Video Technology, Peking University, Beijing, China
[2] Singapore University of Technology and Design, Singapore
[3] National University of Singapore, Singapore [4]Peng Cheng Laboratory, Shenzhen, China
{yongxingdai, lingyu}@pku.edu.cn, lixiaotong@stu.pku.edu.cn
jun_liu@sutd.edu.sg, zekuntong@u.nus.edu

Abstract

Domain generalizable (DG) person re-identification (ReID) is a challenging problem because we cannot access any unseen target domain data during training. Almost all the existing DG ReID methods follow the same pipeline where they use a hybrid dataset from multiple source domains for training, and then directly apply the trained model to the unseen target domains for testing. These methods often neglect individual source domains' discriminative characteristics and their relevances w.r.t. the unseen target domains, though both of which can be leveraged to help the model's generalization. To handle the above two issues, we propose a novel method called the relevance-aware mixture of experts (RaMoE), using an effective voting-based mixture mechanism to dynamically leverage source domains' diverse characteristics to improve the model's generalization. Specifically, we propose a decorrelation loss to make the source domain networks (experts) keep the diversity and discriminability of individual domains' characteristics. Besides, we design a voting network to adaptively integrate all the experts' features into the more generalizable aggregated features with domain relevance. Considering the target domains' invisibility during training, we propose a novel learning-to-learn algorithm combined with our relation alignment loss to update the voting network. Extensive experiments demonstrate that our proposed RaMoE outperforms the state-of-the-art methods.



(a) Conventional DG ReID pipeline

(b) Our method

Figure 1. Differences between our method and the conventional DG ReID pipeline. (a) Conventional DG ReID methods generally train a single model on the hybrid dataset from multi-source domains and then apply the trained model to the unseen target domain for testing, which neglects individual domains' discriminative characteristics and target domain's relevance w.r.t. source domains. (b) Our method leverages the complementary information provided by all the source domain networks (also termed as "domain experts"). In testing, we integrate features obtained by source domain experts into an adaptive voting process based on the unseen target domain's relevance w.r.t. source domains.

1. Introduction

In recent years, the research on person re-identification (ReID) has been appealing to academia and industry. The goal of ReID is to identify a person across different camera views. Many works on fully supervised ReID [48, 27, 8] have achieved quite promising performances when training and testing under the same domain (dataset). However, when applying these well-trained ReID models to other domains, the performance often drops significantly because of the domain biases [52]. To tackle this problem, some researchers have studied unsupervised domain adaptation (UDA) methods [65, 15, 57, 16, 6, 10], which utilize the unlabeled target data to finetune and adapt the source-trained model to the target domain. However, existing UDA ReID methods are often not powerful enough to deal with practical application scenarios, because it is sometimes hard to collect target domain training data and time-consuming to finetune the model on these unlabeled samples. As a result, domain generalizable (DG) ReID [45, 29, 30] has been appealing to researchers recently. Generally, DG ReID methods utilize labeled data from multiple source domains to learn a generalizable model for new unseen target domains,

[*] Corresponding Author.

without using any target domain data for training. To obtain more generalizable models for unseen target domains, we are devoted to the problem of DG ReID in this paper.

Almost all the existing DG ReID methods [45, 29, 30] follow the same pipeline, where they collect all source domain data into a hybrid dataset and train a single model on it, as shown in Fig. 1 (a). During testing, they usually use the same well-trained model to extract features for any unseen target domain. However, there can be two potential problems in such a pipeline: (1) They learn a common feature space for different domains, which may neglect individual domains' discriminative characteristics. Such diverse domain-specific characteristics have been shown to be able to provide complementary information for better generalization on target domains, as mentioned in [13, 21, 67]. (2) Conventional DG ReID methods often ignore the specific target domain's inherent relevance w.r.t. different source domains. They are difficult to generalize the model to the unseen target domain because the model trained on the more relevant source domains can provide more discriminative and meaningful information than those less relevant domains. However, such relevance is often not explicitly considered by existing works [45, 29, 30].

Recently, works [44, 21] on the mixture of experts [28] (MoE) show that MoE can improve the overall model's capability by mixing multiple networks (*i.e.,* leveraging expterts' complementary information) with a voting procedure. Inspired by this, we propose a novel approach called Relevance-aware Mixture of Experts (RaMoE), as shown in Fig. 1 (b), to handle the above two issues (*i.e.,* complementary information and domain relevance). We argue that, instead of learning a single model on the hybrid domains, we can train a domain-specific network (domain expert) for each source domain to exploit individual domains' discriminative and powerful characteristics. Thus, these domain experts' mixture can keep source domains' diversity and provide rich complementary information, improving the generalization on target domains. Subsequently, we propose an adaptive voting network to calculate the unseen target domain's relevance w.r.t. all source domains. Based on the domain relevance, we can adaptively integrate those source experts' features into the aggregated features by voting. The voting network will assign the more relevant domain experts with higher weights. Thus, those more relevant experts will provide more complementary information to improve the aggregated features' generalizability on the target domain.

Specifically, in our RaMoE method, we propose a decorrelation loss to encourage source domain experts to keep their domains' diverse characteristics, and thus they can provide complementary and discriminative information. Such a decorrelation loss is implemented by minimizing the correlation among the source domain experts because the lower correlation among experts will bring about more complementary information, as mentioned in [2, 41]. Because the target domain is totally unseen during training in DG ReID, it is challenging for the adaptive voting network to well learn the target domain's correct relevance w.r.t. source domains. Inspired by meta-learning (learning-to-learn) that can improve the model's generalization [32, 12, 23] for the unseen target domains in an episodic training paradigm, we propose a novel learning-to-learn algorithm to learn our adaptive voting network. At the beginning of each episodic training iteration, we randomly split source domains into the meta-train (simulated "source domains") and the meta-test (simulated "unseen target domains") to simulate the adaptive voting procedure for the unseen target domain. During each episodic training iteration, the meta-test first obtains the relevance w.r.t. the meta-train using the adaptive voting network. The meta-test can then get two kinds of features: one is the features extracted by the meta-test domain expert, and the other is the aggregated features integrated from multiple meta-train domain experts with the relevance. We propose the relation alignment loss to push the aggregated features to be as discriminative as the features extracted by the meta-test expert. As a result, our RaMoE method can generate very discriminative and generalizable aggregated features for the unseen target domains by adaptively integrating diverse domain experts with the domain relevance.

Our major contributions can be summarized as follows: (1) We propose a novel RaMoE method to tackle the problem of DG ReID by exploiting source domains' complementary information and their relevance w.r.t. the unseen target domain. (2) We propose the decorrelation loss to keep source domains' diversity and encourage source domain experts to provide more complementary and discriminative information. (3) To make the model more generalizable to target domains, we propose a voting network to adaptively integrate source domain experts' features into the aggregated features. Specially, the adaptive voting network is updated with the relation alignment loss in a novel learning-to-learn way. (4) Extensive experiments demonstrate that our method outperforms state-of-the-art DG ReID approaches by a large margin.

To the best of our knowledge, this is the first work that treats DG ReID as a novel mixture-of-experts paradigm via an effective voting-based mixture mechanism.

2. Related Work

Person Re-Identification. Deep supervised person ReID has made great progress in recent years, including but not limited to deep metric learning [25, 9, 7, 47], part-based methods [48, 31, 46, 22], and attention network learning [4, 5, 59]. To handle the problem of domain biases [11, 52] in ReID, researchers proposed unsupervised domain adaptation (UDA) methods [65, 15, 57, 16, 6, 10, 17, 60, 61].

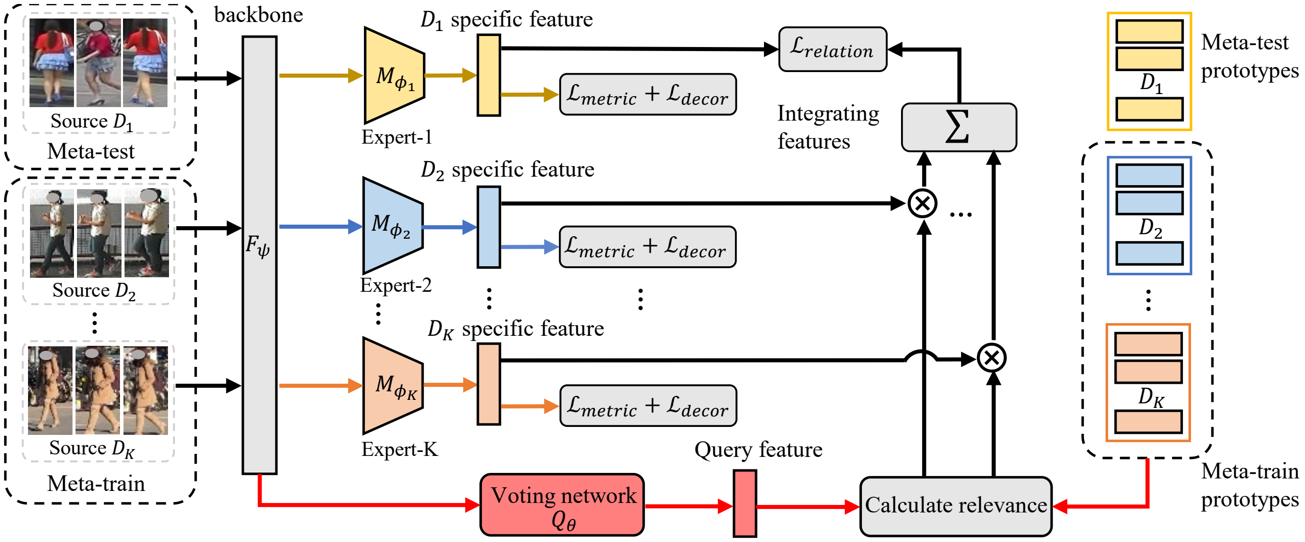


Figure 2. Illustration of our method. The k-th branch network serves as the expert of the domain D_k, and it is learned with the metric loss $\mathcal{L}_{\text{metric}}$ and the decorrelation loss $\mathcal{L}_{\text{decor}}$. We use the learning-to-learn algorithm combined with the relation alignment loss $\mathcal{L}_{relation}$ to update the voting network. At each episodic training iteration, we split K domains into the meta-test (*e.g.,* D_1) and meta-train (*e.g.,* $D_2, ..., D_K$). A meta-test image can obtain K features (one feature from its own domain expert and $K-1$ features from meta-train domain experts), together with a query feature obtained by the voting network. The meta-test domain's relevance w.r.t. the meta-train can be obtained by calculating the mean similarity between the meta-test query feature and the meta-train prototypes. We can obtain the weighted aggregated feature by adaptively integrating meta-train experts' features with the relevance. "$\bigotimes$" is the operation of weighting features with the domain relevance. The relation alignment loss is proposed to push the weighted aggregated feature as discriminative as the meta-test domain-specific feature. "$\sum$" is the features' operation: concatenation or element-wise summation.

Very recently, researchers started to study the topic of domain generalization (DG) in ReID [45, 29, 30], which learns the generalizable ReID models on multi-source domains without using any target training data, and tests on unseen target domains. Song *et al.* [45] proposed the problem of domain generalization in ReID and designed the Domain-Invariant Mapping Network combined with a memory bank to learn domain-invariant features. Jia *et al.* [29] utilized Instance Normalization [49] to learn a more generalizable model. Jin *et al.* [30] proposed Style Normalization and Restitution modules to disentangle the identity-relevant and identity-irrelevant features. Different from all the above DG ReID works, we propose a novel RaMoE method by utilizing individual source domains' diverse and discriminative characteristics and the unseen target domain's relevance w.r.t. source domains in order to adaptively improve the model's generalization on unseen target domains.

Domain Generalization. The goal of general DG is to improve the model generalization in an arbitrary domain for image classification by training from multi-source domains. Existing DG methods can be mainly categorized into three aspects. (1) Learning domain-invariant features [39, 18, 34]: These methods assume that minimizing the domain discrepancy between multi-source domains can help learn domain-invariant features which are robust for unseen target domains. (2) Augmenting source data [43, 50, 3, 67, 66]: These methods augment the source domain data to increase the domain diversity, thus the source-trained model will be more robust to unseen target domains. (3) Optimizing with meta-learning [32, 33, 12]: These methods adopted the episodic training paradigm to split the source domains into meta-train and meta-test to simulate the domain bias, so as to improve the model generalization. The above general DG methods mainly focus on image classification where the target domains and the source domains share the same label space. Thus, these methods can not be directly applied to the task of DG ReID, since in ReID, the identities/classes of the target domains are usually totally different from source domains.

Mixture of Experts. Jacobs *et al.* [28] first introduced the mixture of experts (MoE). MoE aims to learn a system composed of many separated networks (experts), where each expert learns to handle a subset of the whole dataset. Recently, deep MoE methods have shown their superiority in image recognition [1, 20, 51], machine translation [44], scene parsing [14] and so on. Unlike these works, we design a learnable voting network that can be updated with a novel meta-learning algorithm. By integrating all the experts using our designed voting network, we can well leverage the complementary information of those relevant domains' experts to improve the features' generalization in DG ReID.

3. Methodology

In this work, we aim to train a group of experts that are capable of learning discriminative features from their individual domains. When facing an unseen target domain, the mixture of these domain experts can be trained to vote based on their relevances w.r.t. the target domain. By adaptively integrating all the source experts' features into aggregated features with the relevance, our RaMoE can achieve an optimal generalization performance on the target domain.

3.1. Overview

The pipeline of our proposed RaMoE is illustrated in Fig. 2. During training, we can access K source domains' labeled datasets $\boldsymbol{D} = \{D_k\}_{k=1}^{K}$, where $D_k = \{(x_n^k, y_n^k)\}_{n=1}^{N_k}$, (x_n^k, y_n^k) is a labeled sample and N_k is the number of labeled images in the k-th domain. After the backbone F_ψ (*e.g.,* ResNet50), we design K branch networks (termed as "source domain experts") $\{M_{\phi_k}\}_{k=1}^{K}$, and a voting network Q_θ. The metric loss $\mathcal{L}_{\text{metric}}$ makes each expert focus on learning its domain-specific features. The decorrelation loss $\mathcal{L}_{\text{decor}}$ is used to keep source domains' diverse characteristics and encourage all the domain experts to provide complementary information. We use the k-th domain's class centers $C_k = \{c_l^k\}_{l=1}^{L_k}$ as the prototypes to represent the k-th domain's characteristics, where L_k is the number of person identities in the k-th domain.

We propose a novel meta-learning algorithm combined with the relation alignment loss $\mathcal{L}_{\text{relation}}$ to update the voting network. K source domains are randomly split into a meta-test domain and $K-1$ meta-train domains at each episodic training iteration. For a meta-test image, it can obtain $K+1$ features, including (1) a feature extracted by the meta-test expert, (2) $K-1$ features extracted by the meta-train experts, (3) a query feature extracted by the voting network. The meta-test domain's relevance w.r.t. the meta-train domains can be calculated by the mean similarity between the query feature and the meta-train domains' prototypes. We can obtain the weighted aggregated feature by integrating $K-1$ meta-train expert features based on their relevance. The relation alignment loss is proposed to push the weighted aggregated feature to be as discriminative as the meta-test feature.

3.2. Optimizing Domain-specific Experts

As mentioned in [13, 21, 67], exploiting the complementary information of discriminative experts helps improve the overall model's generalization on target domains. Thus, the domain experts should satisfy two properties: discriminability and complementarity. We use the metric loss to improve every domain-specific expert's discriminability. Similar to [41], we mutually reduce all the domain experts' correlation to improve the complementarity among them. Specifically, we propose a decorrelation loss to decorrelate all these domain experts' features.

Metric Loss. Similar to [38], we use the classification loss $\mathcal{L}_{\text{cls}}$, triplet loss [25] $\mathcal{L}_{\text{tri}}$, and center loss [54] $\mathcal{L}_{\text{cent}}$ to optimize K domain-specific experts $\{M_{\phi_k}\}_{k=1}^{K}$, the domain-specific prototypes $\{C_k\}_{k=1}^{K}$, and the backbone network F_ψ. We combine the above metric losses as:

$$\mathcal{L}_{\text{metric}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{tri}} + \lambda \mathcal{L}_{\text{cent}}, \tag{1}$$

where λ (set as 5×10^{-4}) is the weighting hyper-parameter.

Decorrelation Loss. For an image x_n^k (where $n = 1, 2, ..., N_k$) from the k-th domain, we use all the experts to extract K features $\{m_n^j\}_{j=1}^{K}$ that are characterized by individual domains, as shown in Fig. 2. To improve the aggregated features' generalization, we encourage these experts to provide more complementary and discriminative information. Specifically, we propose the decorrelation loss by reducing the correlation among different domain experts. We formulate the decorrelation loss as follows:

$$\mathcal{L}_{\text{decor}} = \frac{1}{N_k} \sum_{n=1}^{N_k} (\frac{1}{K-1} \sum_{j \neq k} ||m_n^k \odot m_n^j||), \tag{2}$$

where features $\{m_n^j\}_{j=1}^{K}$ are all L2-normalized, $\odot$ means the point-wise product and $\|\cdot\|$ is the L2-norm of a vector.

We combine Eq. (1) (2) into the domain loss by:

$$\mathcal{L}_{\text{domain}} = \mathcal{L}_{\text{metric}} + \mathcal{L}_{\text{decor}}. \tag{3}$$

Thus, by alternating k from 1 to K, we can obtain a group of representative and complementary domain experts.

3.3. Optimizing the Voting Network

To make the model more generalizable to the unseen target domain, we leverage the specific target domain's relevance w.r.t. all source domains. Specifically, we propose a voting network to calculate the domain relevance adaptively. By integrating all the source domain experts' features into a weighted aggregated feature with relevance, we can achieve more generalizable features for an unseen target domain during testing. Because the target domain data is unavailable during training, we propose a learning-to-learn algorithm to simulate integrating multi-source experts' features with the relevance. The voting network can be updated with a relation alignment loss introduced below. Thus, we can learn a generalizable voting network for an unseen target domain, integrating multi-source experts' features adaptively. Specifically, we split the K source domains into meta-train (simulated "source domains") $\boldsymbol{D}_s$ including $K-1$ domains, and the meta-test (simulated "the unseen target domain") $\boldsymbol{D}_u$ including the remaining domain, at every episodic training iteration.

Relation Alignment Loss. As mentioned before, for a k-th domain's image x_n^k (where $n = 1, 2, ..., N_k$), we can obtain K features $\{m_n^j\}_{j=1}^{K}$ extracted by K experts $\{M_{\phi_j}(\cdot)\}_{j=1}^{K}$, and a query feature q_n^k extracted by the voting network $Q_\theta(\cdot)$, as illustrated in Fig. 2.

We use the query feature q_n^k to calculate the domain relevance score of the k-th domain's image x_n^k w.r.t. the j-th domain ($j \neq k$) by:

$$s_n^j = \frac{1}{L_j} \sum_{l=1}^{L_j} \langle q_n^k, c_l^j \rangle, \tag{4}$$

where $\langle q_n^k, c_l^j \rangle$ is the inner product between the query feature q_n^k and the l-th class prototype c_l^j (where $l = 1, 2, ..., L_j$) in the j-th domain. Both q_n^k and c_l^j are L2-normalized. As a result, we can get the relevance set $\{s_n^j\}_{j=1,j \neq k}^K$ of the image x_n^k w.r.t. all other $K-1$ domains. Thus, for a k-th domain image x_n^k, we can then integrate other $K-1$ irrelevant experts' features $\{m_n^j\}_{j=1,j \neq k}^K$ into the weighted aggregated feature v_n with the relevance s_n^j by:

$$v_n = \sum_{j \neq k} \sigma(s_n^j) \cdot m_n^j, \tag{5}$$

where $\sigma(\cdot)$ is the non-linear function (*e.g.,* sigmoid or softmax) to normalize the relevance between 0 and 1.

Softmax-triplet function [16, 56] has been shown to be a powerful tool to measure the metric relationship in the feature space (*i.e.,* inter-sample discriminability). Thus we use it to measure the metric relationship of the weighted aggregated feature v_n as below:

$$R(v_n) = \frac{\exp(\|v_n - v_n^+\|)}{\exp(\|v_n - v_n^+\|) + \exp(\|v_n - v_n^-\|)}, \tag{6}$$

where $R(\cdot) \in [0, 1]$, $\|\cdot\|$ is the L2-norm of a vector, and v_n^+ and v_n^- are the selected features of the hardest positive and negative samples within a mini-batch. Similarly, for the k-th expert's feature m_n^k we can also use Eq. (6) obtain the metric relationship $R(m_n^k)$.

Compared with other $K-1$ domain experts, the k-th domain expert should be able to generate more discriminative feature for the sample x_n^k, while such metric relationship $R(m_n^k)$ reflects the k-th domain-specific discriminative characteristics. Thus, we push the weighted aggregated feature v_n to be as discriminative as the k-th domain-specific feature m_n^k, and meanwhile, enable the weighted aggregated feature to be characterized by the k-th domain, we propose the relation alignment loss below:

$$\mathcal{L}_{\text{relation}} = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathcal{L}_{bce}(R(v_n), R(m_n^k)), \tag{7}$$

where $\mathcal{L}_{\text{bce}}$ is the binary cross-entropy loss. By minimizing Eq. (7), the voting network is pushed to learn to produce reliable relevance scores. Thus, the model can learn powerful generalization capabilities for unseen target domains, by exploiting how to integrate source domains.

Meta Optimizing. Since we cannot access the unseen target domain samples, we design a meta-learning scheme to optimize the above losses. At the **meta-training** stage, we use the meta-train $\boldsymbol{D}_s$ to compute the domain loss with Eq. (3) and the relation alignment loss with Eq. (7) as:

$$\mathcal{L}^s = \mathcal{L}_{\text{domain}}^s(\boldsymbol{D}_s; \psi, \phi_s, \boldsymbol{C}_s) + \mathcal{L}_{\text{relation}}^s(\boldsymbol{D}_s; \psi, \phi_s, \boldsymbol{C}_s, \theta), \tag{8}$$

Algorithm 1: Training Procedure of RaMoE

```
Input: Source domains $\boldsymbol{D} = \{D_k\}_{k=1}^K$; Learning rate
       hyperparameters $\alpha, \beta, \gamma$; Balance hyperparameter $\eta$;
       MaxIters; MaxEpochs.
Output: Backbone feature extractor $F_\psi$; Domain-specific
        experts $\{M_{\phi_k}\}_{k=1}^K$; Prototypes $\{C_k\}_{k=1}^K$; Voting
        network $Q_\theta$.
// For simplicity, we denote $\mathcal{L}_{\text{domain}}$ and $\mathcal{L}_{\text{relation}}$ as $\mathcal{L}_d$ and
  $\mathcal{L}_r$ respectively.
for $epoch = 1$ to MaxEpochs do
    for $iter = 1$ to MaxIters do
        Sample $K-1$ domains as meta-train $\boldsymbol{D}_s$ and the
          remaining as meta-test $\boldsymbol{D}_u$;
        Meta-training:
        Compute losses for $\boldsymbol{D}_s$: $\mathcal{L}^s = \mathcal{L}_d^s + L_r^s(\theta)$;
        Update the voting network parameters by:
          $\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_r^s(\theta)$;
        Meta-testing:
        Compute losses for $\boldsymbol{D}_u$: $\mathcal{L}^u = \mathcal{L}_d^u + \mathcal{L}_r^u(\theta')$;
        Optimizing:
        $\psi \leftarrow \psi - \beta \nabla_\psi (\mathcal{L}_d^s + \mathcal{L}_d^u)$;
        $(\phi_s, \boldsymbol{C}_s) \leftarrow (\phi_s, \boldsymbol{C}_s) - \beta \nabla_{\phi_s, \boldsymbol{C}_s} \mathcal{L}_d^s$;
        $(\phi_u, \boldsymbol{C}_u) \leftarrow (\phi_u, \boldsymbol{C}_u) - \beta \nabla_{\phi_u, \boldsymbol{C}_u} \mathcal{L}_d^u$;
        Meta-optimizing
        $\theta \leftarrow \theta - \gamma((1-\eta) \nabla_\theta \mathcal{L}_r^s(\theta) + \eta \nabla_\theta \mathcal{L}_r^u(\theta'))$;
    end
end
```

where ψ is the parameter of the backbone, ϕ_s is the parameter of the domain-specific experts of $\boldsymbol{D}_s$, $\boldsymbol{C}_s$ is the prototypes set of $\boldsymbol{D}_s$, and θ is the parameter of the voting network. Similar to [54], prototypes can be updated with the center loss in Eq. (1). Next, the updated parameters of the voting network is obtained by: $\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{relation}}^s(\theta)$, where α is the learning rate hyper-parameter. At the **meta-testing** stage, we use the meta-test $\boldsymbol{D}_u$ to compute the domain loss and relation alignment loss with Eq. (3) (7), which is formulated as follows:

$$\mathcal{L}^u = \mathcal{L}_{\text{domain}}^u(\boldsymbol{D}_u; \psi, \phi_u, \boldsymbol{C}_u) + \mathcal{L}_{\text{relation}}^u(\boldsymbol{D}_u; \psi, \phi_u, \boldsymbol{C}_u, \theta'), \tag{9}$$

where ϕ_u is the parameter of the $\boldsymbol{D}_u$ expert, $\boldsymbol{C}_u$ is the prototypes set of $\boldsymbol{D}_u$, and θ' is the updated parameter with Eq. (8). At the **meta-optimizing** stage, we optimize the voting network with the second-order gradient as follows:

$$\theta \leftarrow \theta - \gamma((1-\eta) \nabla_\theta \mathcal{L}_{\text{relation}}^s(\theta) + \eta \nabla_\theta \mathcal{L}_{\text{relation}}^u(\theta')), \tag{10}$$

where γ is the learning rate and η (set as 0.5) is the hyper-parameter to balance the gradient of meta-train and meta-test. The overall training procedure is shown in Alg. 1.

3.4. Testing Procedure

During testing, for the unseen target domain dataset consisting of N samples $\{x_n\}_{n=1}^N$, we use Eq. (4) to obtain the relevance of each target sample x_n w.r.t. all K source domains, *i.e.,* $\{s_n^k\}_{k=1}^K$. Then, we can obtain the relevance of the unseen target domain w.r.t. the k-th source domain by $s^k = \frac{1}{N}\sum_{n=1}^{N} s_n^k$. Each target sample x_n can achieve K features $\{m_n^k\}_{k=1}^{K}$ using K domain experts. Similar to Eq. (5), we adaptively integrate all K source domains' features with the relevance $\{s_n^k\}_{k=1}^{K}$ by:

$$v_n = \sum_{k=1}^{K} \sigma(s^k) \cdot m_n^k \ , \tag{11}$$

where the weighted aggregated features $\{v_n\}_{n=1}^{N}$ are all L2-normalized for evaluating.

4. Experiments

4.1. Implementation Details

We use ResNet50 [24] pretrained on ImageNet as our backbone. Similar to [38], the last residual layer's stride size is set as 1. After the global pooling layer we add an Embedding layer (*i.e.,* FC: 2048d→512d) followed by batch normalization (BN) to get the ReID feature. The identity classifier (Classifier) followed by softmax function is added after BN to optimize with the classification loss. The above network is the structure of our ***Baseline***. For efficiency, in our method, we make all the source domains share the same backbone and add a branch network (expert) for each source domain. Specifically, the structure of every domain expert is Embedding→BN→Classifier. The voting network can be easily implemented with FC→ReLU→BN, where FC is 2048d→512d. We resize the person image size to 256 × 128. For data augmentation, we perform random cropping, random flipping, and color jittering. Similar to [30], we discard random erasing (REA) because REA will degenerate the cross-domain ReID performance [38]. The batch size is set to 64, including 16 identities and four images per identity. For our ***Baseline***, we combine all the source domains into a hybrid dataset and only use the metric loss $\mathcal{L}_{\text{metric}}$ for training. In our **RaMoE** method, we sample each source domain evenly at every training iteration. We optimize the model with the Adam optimizer. We train the model for 120 epochs and use the warmup strategy in the first ten epochs. The learning rate (*i.e.,* α, β, γ in Alg. 1) is initialized as 3.5×10^{-4} and divided by 10 at the 40th and 70th epochs respectively. We conduct all the experiments with PyTorch and train the model on four 1080Ti GPUs. The training and testing are efficient in our multi-head RaMoE method where the training and inference time of each batch are 0.708s and 0.312s respectively (batch size is 64).

4.2. Datasets and Evaluation Settings.

Datasets and Evaluation Metrics. Following the previous works [45, 29, 30] on DG ReID, we conduct our experiments on the public ReID or Pearson-Search datasets, including Market1501 [62], DukeMTMC-reID [63], CUHK02 [35], CUHK03 [36], MSMT17 [52], CUHK-SYSU [55], and four small ReID datasets including PRID [26], GRID [37], VIPeR [19], and iLIDs [53]. For CUHK03, we use the "labelled" dataset for training and adopt the protocol used in [64] for testing. For simplicity, in the next sections we denote Market1501 as M, DukeMTMC-reID as D, CUHK02 as C2, CUHK03 as C3, MSMT17 as MT, and CUHK-SYSU as CS. We use the mean average precision (mAP) and Cumulative Matching Characteristics (CMC) for evaluation.

Table 1. Different evaluation protocols. The leave-one-out setting for M+D+C3+MT means selecting one domain for testing and the remaining three domains for training.

Setting	Training Data	Testing Data
Protocol-1	M+D+C2+C3+CS	PRID, GRID, VIPeR, iLIDs
Protocol-2	M+D+C3+MT	
Protocol-3	Leave-one-out for M+D+C3+MT	

Evaluation Protocols. There exist two evaluation protocols for DG ReID, as shown in Tab. 1. Under the setting of Protoco1-1 [45], all the images in these datasets M+D+C2+C3+CS (including the training and testing sets) are used for training. Four small ReID datasets (*i.e.,* PRID, GRID, VIPeR, and iLIDs) are tested respectively, where the final performances of these small ReID datasets are evaluated on the average of 10 repeated random splits of gallery and probe sets. Under Protocol-2 [30], all the images in M+D+C3+MT (including the training and testing sets) are used for training and the testing sets are the same as Protocol-1. However, two disadvantages may lie in Protocol-1 and Protocol-2: (1) Compared with the existing ReID datasets, the number of images per identity in the CS dataset is much smaller, which will limit the learning of discriminative ReID features. (2) The images' quality of the four small ReID datasets is low. The small datasets' performances can not correctly evaluate the model's generalizability in real scenarios, where the latter needs to be evaluated on large-scale datasets. As a result, we set a new protocol (*i.e.,* Protocol-3 in Tab. 1) of the leave-one-out setting for the existing large-scale public datasets M+D+C3+MT. Specifically, the leave-one-out setting of M+D+C3+MT is selecting one domain from M+D+C3+MT for testing (only the testing set in this domain) and all the remaining domains for training (including the training and testing sets).

4.3. Comparison with the State-of-the-Arts

Our proposed RaMoE can outperform the state-of-the-arts methods by a large margin in the task of Domain Generalization (DG) ReID, as shown in Tab. 2. The ***Baseline*** method is training on the hybrid dataset including all source domains with only the metric loss $\mathcal{L}_{\text{metric}}$.

Comparison with DG ReID methods under the Protocol-1 and Protocol-2. We compare our method with the existing DG ReID methods under two different evaluation protocols. All the other methods directly apply the model trained on source domains to the unseen target do-

Table 2. Comparison with state-of-the-arts methods in DG ReID under the setting of protocol-1 and protocol-2. We report the performances of the methods marked by “ * ” from [45]. The best results are highlighted with bold.

Setting	Method	Reference	Target: PRID		Target: GRID		Target: VIPeR		Target: iLIDs	
			mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Protocol-1	Agg_Align* [58]	arXiv 2017	25.5	17.2	24.7	15.9	52.9	42.8	74.7	63.8
	Reptile* [40]	arXiv 2018	26.9	17.9	23.0	16.2	31.3	22.1	67.1	56.0
	CrossGrad* [43]	ICLR 2018	28.2	18.8	16.0	8.96	30.4	20.9	61.3	49.7
	Agg_PCB* [48]	TPAMI 2019	32.0	21.5	44.7	36.0	45.4	38.1	73.9	66.7
	MLDG* [32]	AAAI 2018	35.4	24.0	23.6	15.8	33.5	23.5	65.2	53.8
	PPA* [42]	CVPR 2018	45.3	31.9	38.0	26.9	54.5	45.1	72.7	64.5
	DIMN* [45]	CVPR 2019	52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2
	SNR [30]	CVPR 2020	66.5	52.1	47.7	40.2	61.3	52.9	89.9	84.1
	Baseline	CVPR 2021	60.4	47.3	49.0	39.4	58.0	49.2	84.0	77.3
	RaMoE (Ours)		**67.3**	**57.7**	**54.2**	**46.8**	**64.6**	**56.6**	**90.2**	**85.0**
Protocol-2	SNR [30]	CVPR 2020	60.0	49.0	41.3	30.4	65.0	55.1	91.9	87.0
	Baseline	CVPR 2021	58.9	47.2	47.7	38.1	63.8	54.7	89.2	84.2
	RaMoE (Ours)		**66.8**	**56.9**	**53.9**	**43.4**	**72.2**	**63.4**	**92.3**	**88.4**

Table 3. Comparisons under the setting of protocol-3.

Target: Market	mAP	Rank-1	Rank-5	Rank-10
Baseline	49.9	75.4	86.9	91.0
RaMoE (Ours)	**56.5**	**82.0**	**91.4**	**94.4**
Target: Duke	mAP	Rank-1	Rank-5	Rank-10
Baseline	49.4	65.8	79.0	83.9
RaMoE (Ours)	**56.9**	**73.6**	**85.3**	**88.4**
Target: CUHK03	mAP	Rank-1	Rank-5	Rank-10
Baseline	32.6	32.9	52.9	63.6
RaMoE (Ours)	**35.5**	**36.6**	**54.3**	**64.6**
Target: MSMT17	mAP	Rank-1	Rank-5	Rank-10
Baseline	9.9	24.5	35.4	40.9
RaMoE (Ours)	**13.5**	**34.1**	**46.0**	**51.8**

main without considering the domain relevance. Compared with them, our RaMoE can outperform them significantly.

Comparison under the Protocol-3. We compare our proposed **RaMoE** with the ***Baseline*** method under the protocol-3 in Tab. 3. The performances on these large-scale ReID datasets have shown our method's superiority in integrating source domains' characteristics adaptively for better domain generalization.

4.4. Ablation Study

Effectiveness of the domain decorrelation. We propose the decorrelation loss $\mathcal{L}_{\text{decor}}$ to encourage source domain experts to keep their diverse and discriminative characteristics. Thus, integrating these experts can provide complementary information to improve the aggregated features' generalization. As shown in Tab. 4, our method outperforms ours w/o decorrelation by 1.3% in Rank-1 on PRID. If learning source experts without the decorrelation loss, the experts will provide less complementary information and thus reduce the generalization of the aggregated features.

Effectiveness of the voting network. The voting network learned with meta-learning can adaptively provide the relevance of the target domain w.r.t. source domains, making those more relevant source domains provide more complementary information to improve the generalization of the weighted aggregated features. As shown in Tab. 4, our method outperforms ours w/o voting (Experts-ensemble) by 1.7%, 1.6%, 1.6%, 0.9% in mAP on PRID, GRID, VIPeR, and iLIDs respectively. Experts-ensemble means that the relevance of the target domain w.r.t. source domains is set 1, and all the experts' features are directly concatenated into the ensemble features. However, our method uses the domain relevance to integrate adaptively. Take the performances on iLIDs as an example, the Expert-M performs worst compared with other three experts (*i.e.,* Expert-D/C3/MT) and the Expert-D performs best. Though directly mixing all these experts (*i.e.,* Experts-ensemble) can bring about great performance gain, the methods w/o voting is inferior to our RaMoE significantly. It can demonstrate that the voting mechanism using the domain relevance can adaptively leverage those more relevant experts' complementary information and alleviate the influence of those less relevant experts.

Can individual domain experts provide complementary information to improve the features' generalization? We can keep all the source domains' diverse and discriminative characteristics using the decorrelation loss. Thus, all the source domain experts are encouraged to provide more complementary information. As shown in Tab. 4, almost all the experts (*i.e.,* Expert-M/D/C3/MT) do not perform very well on different target domains. However, when integrating these experts' features, the aggregated features are superior to those extracted by individual experts. Thus, we can improve the overall features' generalization for unseen target domains by leveraging individual source domains' complementary information.

How to integrate different source domain features? As shown in Tab. 5, we compare different combinations of non-linear functions $\sigma(\cdot)$ and feature integrating types. The results show that the types of the non-linear function $\sigma(\cdot)$ in Eq. (5) will not bring about significant performance fluctuations. When concatenating features obtained by different source domain experts, the performance is better than summing features along with the corresponding dimensions, because the type of concatenating will keep more information

Table 4. We study ablation studies on individual components of our method under the Protocol-2. Voting means learning the voting network with meta-learning by $\mathcal{L}_{\text{relation}}$ and decorrelation means decorrelating source domain experts by $\mathcal{L}_{\text{decor}}$. Expert-M/D/C3/MT means using the feature extracted by Market/Duke/CUHK03/MSMT17 domain expert. Experts-ensemble means concatenating source domain experts' features directly without learning the domain relevance.

Method	Target: PRID		Target: GRID		Target: VIPeR		Target: iLIDs	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
w/o voting (Expert-M)	62.2	53.4	50.4	39.8	66.1	56.9	87.9	82.7
w/o voting (Expert-D)	61.6	51.6	48.4	38.5	67.0	57.5	90.3	85.5
w/o voting (Expert-C3)	62.5	53.7	51.0	41.4	68.5	59.7	88.7	84.0
w/o voting (Expert-MT)	63.6	54.9	49.9	40.0	66.9	57.3	89.0	84.5
w/o voting (Experts-ensemble)	65.1	56.1	52.3	42.2	70.6	61.6	91.4	86.7
w/o decorrelation	66.0	55.6	53.2	42.9	71.3	62.8	91.2	87.0
RaMoE (Ours)	**66.8**	**56.9**	**53.9**	**43.4**	**72.2**	**63.4**	**92.3**	**88.4**

Table 5. Evaluating on different non-linear functions $\sigma(\cdot)$ and feature integrating types under the setting of Protocol-2.

Non-linear $\sigma(\cdot)$		Integrating type		Target: GRID	
softmax	sigmoid	concat	sum	mAP	R1
✓		✓		53.9	43.4
	✓	✓		53.7	43.3
✓			✓	52.3	41.2
	✓		✓	51.9	40.9

Table 6. Evaluation of mAP within source domains.

Method	M	D	C3	MT
Single-source Baseline	81.8	71.6	62.0	46.6
Multi-source Baseline	82.6	74.4	64.3	48.0
RaMoE (Ours)	83.8	74.6	65.6	49.1

about the different feature dimensions. Thus, we choose the combination of "softmax" and "concat" to integrate source domains' features for our method in all the experiments.

4.5. Extension

Evaluation on source domains. We use M, D, C3, and MT as source datasets, where only their training sets are used to train, and their testing sets are only used to test. In Tab. 6, Single-source Baseline means training and testing on the single domain; Multi-source Baseline means training on a hybrid dataset of all domains and testing on each domain separately. Comparing with them, the accuracy of RaMoE on source datasets does not drop but increases.

Extension to the online setting. We can easily extend our testing procedure to a more practical setting where the query set samples are given online, *i.e.,* only the gallery samples are used to calculate the domain relevance in testing. We compare our method with this online setting in Tab. 7 and there is only very negligible performance drop when only using gallery to calculate the domain relevance.

Visualization. As shown in Fig. 3, we visualize the domain relevance between the target domains (*i.e.,* PRID, GRID, VIPeR, and iLIDs) w.r.t. the source domains (*i.e.,* Market, Duke, CUHK03, and MSMT17), where the domain relevance is calculated with the manner mentioned in Sec. 3.4. In Fig. 3, we can see that the unseen target domain's relevance w.r.t. all the source domains are different, and there exist some more relevant source domains for the unseen target domain. For the unseen target domain dataset iLIDs, its style is more similar to MSMT17 and Duke, and thus their relevances are higher than the other two datasets.

Table 7. Evaluation of mAP on different kinds of calculating the domain relevances $\{s^k\}$ in testing under the setting of Protocol-3.

Samples used for calculating $\{s^k\}$	M	D	C3	MT
All samples (Ours)	56.5	56.9	35.5	13.5
Only samples in gallery	56.3	56.8	35.4	13.5

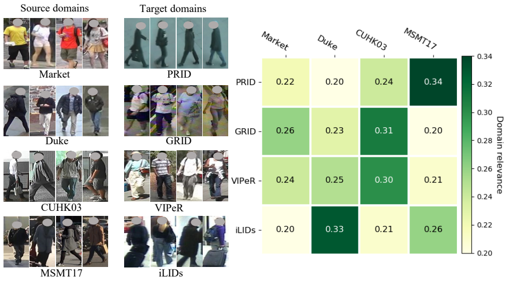


Figure 3. Visualization on the domain relevance.

5. Conclusion

This paper proposes a novel approach called Relevance-aware Mixture of Experts (RaMoE) to tackle the problem of domain generalizable person ReID (DG ReID). By considering both the source domains' individual discriminative characteristics and the relevance of the unseen target domain w.r.t. source domains, we can obtain more generalizable features adaptively for the unseen target domain in DG ReID. Specifically, we propose the decorrelation loss to keep source domains' diverse and discriminative characteristics. Thus, these experts can provide more complementary information to improve the aggregated features' generalization. To obtain more accurate domain relevance of the unseen target domain w.r.t. source domains, we propose the voting network learned with the relation alignment loss in a meta-learning way. Extensive experiments show the effectiveness of our proposed RaMoE method.

Acknowledgements: 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, and was partially supported by SUTD Project PIE-SGP-A12020-02.

References

[1] Karim Ahmed, Mohammad Haris Baig, and Lorenzo Torresani. Network of experts for large-scale image categorization. In *ECCV*, pages 516–532, 2016. 3

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 2

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, pages 2229–2238, 2019. 3

[4] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, pages 371–381, 2019. 2

[5] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *ICCV*, pages 9637–9646, 2019. 2

[6] Guangyi Chen, Yuhao Lu, Jiwen Lu, and Jie Zhou. Deep credible metric learning for unsupervised domain adaptation person re-identification. In *ECCV*, pages 643–659, 2020. 1, 2

[7] Guangyi Chen, Tianren Zhang, Jiwen Lu, and Jie Zhou. Deep meta metric learning. In *ICCV*, pages 9547–9556, 2019. 2

[8] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8351–8361, 2019. 1

[9] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017. 2

[10] Yongxing Dai, Jun Liu, Yan Bai, Zekun Tong, and Ling-Yu Duan. Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. *arXiv preprint arXiv:2012.13689*, 2020. 1, 2

[11] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, pages 994–1003, 2018. 2

[12] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, pages 6450–6461, 2019. 2, 3

[13] Antonio D'Innocente and Barbara Caputo. Domain generalization with domain-specific aggregation modules. In *German Conference on Pattern Recognition*, pages 187–198. Springer, 2018. 2, 4

[14] Huan Fu, Mingming Gong, Chaohui Wang, and Dacheng Tao. Moe-spnet: A mixture-of-experts scene parsing network. *Elsevier PR*, 84:226–236, 2018. 3

[15] 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*, pages 6112–6121, 2019. 1, 2

[16] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 1, 2, 5

[17] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020. 2

[18] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pages 2551–2559, 2015. 3

[19] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008. 6

[20] Sam Gross, Marc'Aurelio Ranzato, and Arthur Szlam. Hard mixtures of experts for large scale weakly supervised vision. In *CVPR*, pages 6865–6873, 2017. 3

[21] Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *EMNLP*, pages 4694–4703, 2018. 2, 4

[22] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, pages 3642–3651, 2019. 2

[23] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *CVPR*, pages 6163–6172, 2020. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[25] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 4

[26] 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. Springer, 2011. 6

[27] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, pages 9317–9326, 2019. 1

[28] 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, 3

[29] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. *BMVC*, 2019. 1, 2, 3, 6

[30] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020. 1, 2, 3, 6, 7

[31] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, pages 384–393, 2017. 2

[32] Da Li, Yongxin Yang, Yi-Zhe Song, and TM Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, pages 3490–3497, 2018. 2, 3, 7

[33] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, pages 1446–1455, 2019. 3

[34] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018. 3

[35] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601, 2013. 6

[36] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 6

[37] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *Springer IJCV*, 90(1):106–129, 2010. 6

[38] 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 TMM*, 2019. 4, 6

[39] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18, 2013. 3

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

[41] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE TPAMI*, 2018. 2, 4

[42] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, pages 7229–7238, 2018. 7

[43] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018. 3, 7

[44] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 2, 3

[45] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, pages 719–728, 2019. 1, 2, 3, 6, 7

[46] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, pages 402–419, 2018. 2

[47] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. 2

[48] Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE TPAMI*, 2019. 1, 2, 7

[49] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 3

[50] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, pages 5334–5344, 2018. 3

[51] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *UAI*, pages 552–562. PMLR, 2020. 3

[52] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 1, 2, 6

[53] Zheng Wei-Shi, Gong Shaogang, and Xiang Tao. Associating groups of people. In *BMVC*, pages 23–1, 2009. 6

[54] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016. 4, 5

[55] 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. 6

[56] Han-Jia Ye, Su Lu, and De-Chuan Zhan. Distilling cross-task knowledge via relationship matching. In *CVPR*, pages 12396–12405, 2020. 5

[57] 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*, pages 9021–9030, 2020. 1, 2

[58] 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

[59] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 2

[60] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhan, and Zheng-Jun Zha. Exploiting sample uncertainty for domain adaptive person re-identification. In *AAAI*, 2021. 2

[61] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *CVPR*, 2021. 2

[62] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 6

[63] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017. 6

[64] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017. 6

[65] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, pages 598–607, 2019. 1, 2

[66] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020. 3

[67] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *arXiv preprint arXiv:2003.07325*, 2020. 2, 3, 4