Mitigate Domain Shift by Primary-Auxiliary Objectives Association for Generalizing Person ReID

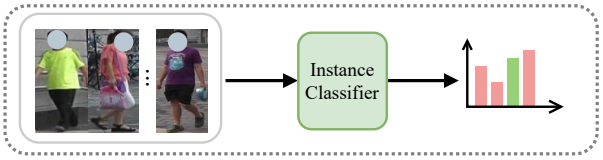
Qilei Li, Shaogang Gong
Queen Mary University of London
{q.li, s.gong}@qmul.ac.uk

Abstract

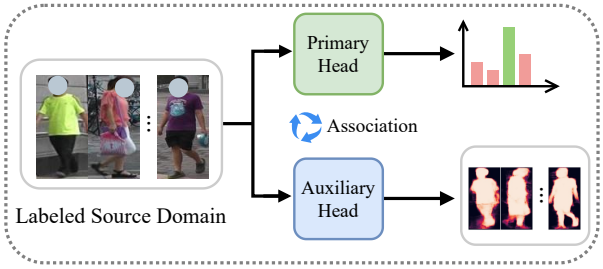
While deep learning has significantly improved ReID model accuracy under the independent and identical distribution (IID) assumption, it has also become clear that such models degrade notably when applied to an unseen novel domain due to unpredictable/unknown domain shift. Contemporary domain generalization (DG) ReID models struggle in learning domain-invariant representation solely through training on an instance classification objective. We consider that a deep learning model is heavily influenced and therefore biased towards domain-specific characteristics, e.g.*, background clutter, scale and viewpoint variations, limiting the generalizability of the learned model, and hypothesize that the pedestrians are domain invariant owning they share the same structural characteristics. To enable the ReID model to be less domain-specific from these pure pedestrians, we introduce a method that guides model learning of the primary ReID instance classification objective by a concurrent auxiliary learning objective on weakly labeled pedestrian saliency detection. To solve the problem of conflicting optimization criteria in the model parameter space between the two learning objectives, we introduce a Primary-Auxiliary Objectives Association (PAOA) mechanism to calibrate the loss gradients of the auxiliary task towards the primary learning task gradients. Benefiting from the harmonious multitask learning design, our model can be extended with the recent test-time diagram to form the PAOA+, which performs on-the-fly optimization against the auxiliary objective in order to maximize the model's generative capacity in the test target domain. Experiments demonstrate the superiority of the proposed PAOA model.*

1. Introduction

Person Re-IDentification (ReID) [18, 21, 40, 45] is a fundamental task which aims to retrieve the same pedestrian across non-overlapping camera views by measuring the distances among representations of all the candidates in a pre-learned discriminative feature space. However, like most deep-learning models, current ReID techniques are built based on an intrinsic assumption of independent and identical distribution (IID) between training and test data. The IID assumption becomes mostly invalid across different domains when training and test data are not from the same environment. As a result, most contemporary ReID models suffer from dramatic degradation when applied to a new domain [4, 25, 34]. Domain Generalization (DG) methods [26, 46, 47], which aim to learn a generalizable model between a source and a target domain have been explored by recent studies to address this problem.



(a) Conventional ReID model training pipeline



(b) The proposed Primary-Auxiliary Objective Association

Figure 1. Comparing a standard Domain Generalization ReID model and the proposed *Primary-Auxiliary Objectives Association* (PAOA) model. A DG model is typically trained by optimizing an instance classification objective, which can suffer from overfitting to domain-specific characteristics, *e.g.*, luminance, background, scale, and viewpoint. The PAOA model considers learning jointly a weakly labeled/supervised auxiliary saliency detection task concurrently with the primary task of the discriminative person ReID. This is achieved by calibrating the gradient of the auxiliary task against that of the primary objective as its reference.

A number of DG ReID methods have been developed to

mitigate performance degradation caused by domain shift between training (source) data and test (target) data. They can be broadly categorized into three main groups: (1) Learning from diversified training samples [1, 16], (2) Aligning the distribution of source domains by data statistics [14, 49, 50], (3) Exploiting meta-learning [4, 5, 42, 47] to mimic source-target distribution discrepancies. The first category confers advantages to a model through the utilization of a diversified training dataset by either image sample augmentation or feature distribution expansion. The second category aims to learn a source-invariant model by aligning the training data, and expecting it to be invariant for the target domain. The third category focuses on simulating the training/testing discrepancy. Despite some performance improvement from these methods, their overall performances across domains remain poor, *e.g.*, the latest SOTA models [4, 42] can only achieve below 20% mAP on the MSMT17 benchmark. This highlights the limitation of overfitting in the current DG ReID models and their inability to learn a more generalizable cross-domain model representation. We consider this is due to the not-insignificant interference of domain-specific contextual scene characteristics such as background, viewpoint, and object distances to a camera (scale), which are identity-irrelevant but can change significantly across different domains. Contemporary DG ReID models are mostly trained by an instance-wise classification objective function, indirectly learning person foreground attention selection (Figure 1(a)). They are sensitive to such domain-specific but identity-irrelevant contextual information, resulting in the misrepresentation of person foreground attention and leading to less discriminative ReID representation. This likely causes notable ReID performance degradation on models trained and deployed in different domains. To mitigate the impact of domain-specific contextual attributes, an intuitive solution is to isolate the pedestrian object to acquire a domain-invariant representation. Several endeavors [11, 29, 48] have been made to guide the person identification network focusing on the pedestrian with the human saliency prior, which can point out the attentive region relevant to the human subject. These methods have certain limitations, either relying on exhaustive manual masking [29] or lacking an appropriate training objective [11, 48] to ensure the accuracy of the generated segmentation mask. Besides this, it is crucial to note that these methods fail to consider the potential worst-case scenario in which the saliency attention prior may be inaccurate, further leading to negative impacts on identification rather than improvement.

In this work, we address this problem by introducing a novel model learning regularization method called *Primary-Auxiliary Objectives Association* (PAOA). Our aim is to minimize domain-specific contextual interference in model learning by focusing more on the domain-invariant person's unique characteristics. This is achieved by introducing the association of learning the primary instance classification objective function with an auxiliary weakly labeled/supervised pedestrian saliency detection objective function, the idea is illustrated in Figure 1(b). Specifically, PAOA is realized in two parts: (1) Additionally train a pedestrian saliency detection head with an auxiliary supervision to assist in focusing the primary ReID discriminative learning task on more domain-invariant feature characteristics. (2) Eliminate the interference attributed to inaccurate saliency labels by calibrating the gradients of the shared feature extractor raised from the weakly-labeled auxiliary learning task towards that of the primary task as a reference when they are in conflict [28]. This association mechanism helps ensure the ReID model learns to attentively focus on generic yet discriminative pedestrian information whilst both learning tasks are harmoniously trained.

Our contributions are: (1) We introduce the idea of optimizing a more domain-generic ReID learning task that emphasizes domain-invariant pedestrian characteristics by associating the ReID instance discriminative learning objective to an auxiliary pedestrian saliency detection objective in a way that does not create conflicts or hinder the effectiveness of primary objective. (2) We formulate a novel regularization called Primary-Auxiliary Objectives Association (PAOA) to implement the proposed association learning. It jointly trains the primary and auxiliary tasks with referenced gradient calibration to solve the conflicting optimization criteria between the two learning objectives, and promote the learning of a more domain-generic ReID model. (3) We further explore the target domain test data characteristics by incorporating the PAOA regularization into a deployment-time model online optimization process. To that end, we formulate a PAOA+ mechanism for on-the-fly target-aware model optimization and show its performance benefit.

2. Related Work

Domain Generalizable ReID (DG ReID) assuming the absence of target domains during training, aims to learn a generalizable model which can extract discriminative representations in any new environment. It's naturally challenging but practical and has attracted increasing attention. Contemporary studies typically fall into three primary classifications: (1) To benefit the model from the diverse training data achieved by augmentation. (2) To align the target domain with the BN statistics calculated over the source domain. (3) To mimic the train/test discrepancy with meta-learning. Despite the improvement obtained by these SOTA models, significant room for improvement remains, as indicated by the low mAP scores, *e.g.*, less than 20% on MSMT17 and less than 40% on CUHK03. This is attributed to the domain-specific interference in the source domain that limits the learning of a domain-invariant model. In this work, we aim

to tackle this issue by guiding the model to focus on the discriminative pedestrian area with the tailored auxiliary task, and propose the PAOA regularization for that end.

Salient Object Detection [3] aims to identify objects or regions that are visually more attentive than the surrounding areas. It has been significantly boosted solely by the rapid development of deep learning. Current detection models are usually trained end-to-end and output a fine-grained saliency map at the pixel level. In this work, we design the auxiliary task with the pedestrian saliency detection objective. Instead of exhaustively labeling the pedestrian area manually as the previous work [29], we propose to use weakly labeled data generated by a trained salient object detection model, to benefit from large-scale training. The recent work GASM [11] shares the similar spirit to ours by employing weakly labelled saliency masks as an additional prior. However, GASM simply trains the saliency detection layers with the classification network while omitting the potential worst-case where the weak label is not accurate and cause potential conflict optimization direction during model training. In contrast, our method focuses on the *association* between instance classification and saliency detection objectives by the proposed referenced gradient calibration mechanism, which promotes the learning of the primary objective while mitigating the conflicts between the primary and auxiliary tasks.

Multitask learning [39] emerges as a solution to learn a single model which is shared across several tasks, so as to achieve greater efficiency than training dedicated models individually for each task. Recent work [37] pointed out that conflicting gradients during multitask learning impede advancement. To break this condition and achieve positive interactions between tasks, they proposed to de-conflict such gradients by altering their directions towards a common orientation. Our model is also constructed in a multitask learning manner, in which the main and the auxiliary tasks are jointly optimized during training. However, the auxiliary task is designed to facilitate the main task therefore it is unsuitable to consider them in the same hierarchy. Instead, we propose referenced gradient calibration by setting the main task as the reference, and calibrating the auxiliary gradient towards it, so as to ensure the auxiliary task can be harmoniously trained alongside the main task, so that it may provide supervision for the primary model objective.

Test-Time model optimization is an emerging paradigm to tackle distribution shifts between training and testing environments. The key idea is to perform post-training model optimization given the test samples during deployment. Several recent works [7, 13, 32, 33] proposed to optimize the model parameters by providing proper supervision, such as batch-norm statistics, entropy minimization, and pseudo-labeling. Another line of work [23, 31] jointly trains additional self-supervised auxiliary tasks, which are subsequently used to guide the model optimization during testing. This does not involve any assumptions about the output and is therefore more generic. It has also been applied to ReID [9] by considering self-supervised learning tasks for updating BN statistics. In this work, we formulate PAOA+ by incorporating the proposed PAOA regularization into the deployment-time optimization framework to seek further improvement. With the tailored auxiliary objective as the optimization supervision, PAOA+ effectively exploits the underlying target domain characteristic and exhibits boosted performance on all the benchmarks.

3. Methodology

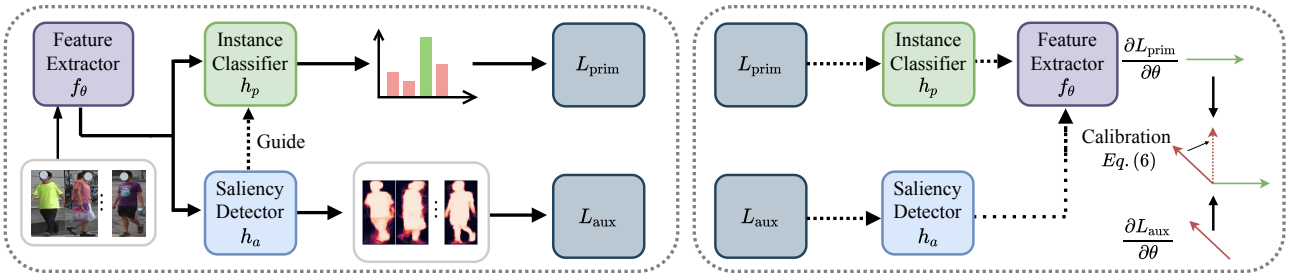
Problem Definition Given a labeled source domain $\mathcal{D}_S = \{(x_i, y_i)\}_{i\in\{1,\cdots,N\}}$ for training, where N is the number of samples, the aim of ReID is to learn a mapping function parameterized by θ that projects a person image x to a high-dimensional feature representation f_θ, with the constraint that features of the same identity have a smaller distance relative to one another. DG ReID is more practical by assuming the non-availability of the target domain during training, and expects the model to be able to extract discriminative feature representations from any target domain. Current models are designed solely with an instance classification objective, that can be confused by negative domain-specific information and fall into a local optimum of the source domain.

3.1. Overview

In this work, we consider the problem of generalizing a ReID model to any new deployment target environment subject to unknown domain bias between the training and the test domains, where there is no labeled training data from the test domain. To that end, we propose a *Primary-Auxiliary Objectives Association* (PAOA) regularization method to enable the model to be more attentive to learning universal identity generative information that is applicable in any domain whilst concurrently maximizing ReID discriminative information from the domain labeled data. Figure 2 shows an overview of PAOA in model training with two associative steps: (1) Guiding the ReID model to focus on discriminative pedestrian information with an additional auxiliary task dedicated to visual saliency detection. (2) Calibrate the gradients of the auxiliary task when it conflicts with the primary instance classification objective. To boost the performance, we build PAOA+ to utilize the available samples in deployment time by minimizing the proposed auxiliary objective, and demonstrate the plug-and-play merit of our design.

3.2. Joint Primary-Auxiliary Objectives Learning

The primary and auxiliary objectives are jointly trained in a multitask learning architecture, which is composed of a



(a) Model Learning: Forward Step (b) Model Learning: Backward Step

Figure 2. Overview of the proposed *Primary-Auxiliary Objectives Association* (PAOA) model. The purpose is to derive generic feature representations by guiding the network to attentively focus on pedestrian information and mitigate the interference of domain-specific knowledge, which is achieved by the PAOA regularization of a primary classification objective and an auxiliary pedestrian saliency detection objective: (a) The auxiliary task is jointly trained to provide hard-coded spatial attention to the pedestrian region. (b) The primary task is used as a reference to calibrate the gradients of the auxiliary objective when they are conflicting.

shared feature extractor f_θ, and two dedicated heads h_p and h_a respectively for the primary and auxiliary tasks.

Primary Objective: Person ReID Learning a strong instance classification network is fundamentally important for training a discriminative ReID model. Given a labeled training set $\mathcal{D} = \{(x_i, y_i^{(p)})\}_{i\in\{1,\cdots,N\}}$, where x_i is a person image and $y_i^{(p)}$ is the corresponding instance category label, the primary instance classification task is trained with a softmax cross-entropy (CE) loss $\mathcal{L}_{\text{id}}$ and a triplet loss $\mathcal{L}_{\text{tri}}$:

$$\mathcal{L}_{\text{id}} = -\sum_{i=1}^{N}\sum_{j=1}^{C} p_i^j \log \hat{p}_i^j, \tag{1}$$

where p_i is one-hot vector activated at $y_i^{(p)}$, and $\hat{p}_i^j$ is the probability for categorized into the jth class that calculated from the classifier. The additional triplet loss constrains the distance between positive (same identity) and negative (different identities) sample pairs, which is formulated as

$$\mathcal{L}_{\text{tri}} = \sum_{i=1}^{N} [d_p - d_n + \alpha]_+, \tag{2}$$

where d_p and d_n respectively denote the Euclidean distances for the positive and negative pairs in feature space. α is the margin that controls the sensitivity and $[s]_+$ is $\max(s, 0)$. The overall loss function for the primary task is as follows:

$$\mathcal{L}_{\text{prim}} = \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{tri}}. \tag{3}$$

Auxiliary Objective: Pedestrian Saliency Detection As illustrated in [31], an auxiliary task closely aligned with the primary task can substantially prompt the learning of the primary objective. Inspired by this, we formulated the auxiliary task as pedestrian saliency detection to perform pixel-level pedestrian localization within the cropped pedestrian bounding boxes. Such an auxiliary task is complementary to the primary task by providing pixel-level hard-coded spatial attention to guide the ReID model to focus on the pedestrian region. Instead of exhaustively manually annotating the pedestrian region, we benefit from the large-scale trained model [41] and perform feed-forward inference to get the weakly labelled samples. Specifically, given a trained saliency model $\mathcal{G}$, we feed the sample to obtain the weak label as $y_i^{(a)} = \mathcal{G}(x_i)$, which is a 2D map to indicate the saliency area. The auxiliary task and it's essentially a regression task in the pixel level. To that end, the auxiliary head h_a is designed as a lightweight module composed of cascaded 2D CNN layers to predict the saliency map. It is optimized by minimizing a conventional $L1$ loss on the predicted salient label $\hat{y}_k^{(a)}$:

$$\mathcal{L}_{\text{aux}} = \sum_{k=1}^{N_k} |y_k^{(a)} - \hat{y}_k^{(a)}|. \tag{4}$$

Joint Multi-task Learning To build a joint multitask learning pipeline, we formulate the overall objective function by combining both L_{prim} and L_{aux} as

$$\mathcal{L}_{\text{train}} = \frac{1}{N}\sum_{1}^{N} \mathcal{L}_{\text{prim}}(x_i, y_i^{(p)}; f_\theta, h_p) + \lambda \mathcal{L}_{\text{aux}}(x_i, y_i^{(a)}; f_\theta, h_a), \tag{5}$$

where λ is the balancing hyperparameter.

Limitation: Despite the auxiliary objective essentially providing hard-coded spatial attention to guide the network being focused on the salient pedestrian object, this pipeline is intrinsically limited. This is due to the inherent noise in the weak label of the auxiliary task that brings a detrimental impact on the primary task and distracts the shared feature extractor from focusing on the pedestrian region. This has

further resulted in a divergent gradient descent direction, reflected by the conflicting gradients. We intuitively visualize the cause of interference in Figure 3. Hence, it becomes necessary to perform a post-operation that resolves the conflicts between the learning objectives.

3.3. Association: Referenced Gradient Calibration

During the model training, the learnable parameter θ of the shared feature extractor f_θ is updated based on two loss gradients: $\boldsymbol{g_p} = \frac{\partial L_{\text{prim}}}{\partial \theta}$ from the primary objective and $\boldsymbol{g_a} = \frac{\partial L_{\text{aux}}}{\partial \theta}$ from the auxiliary objective. However, when $\boldsymbol{g_p}$ and $\boldsymbol{g_a}$ are in conflict as reflected by a negative inner product, *i.e.*, $(\boldsymbol{g_a} \cdot \boldsymbol{g_p}) < 0$, their joint effort cannot provide the network with an informative direction on which to perform the gradient descent to optimize the parameters. Therefore, collectively they bring significant difficulty in model convergence and can even lead to destructive interference [37].

Figure 3. Illustration of the interference to the ReID objective when the weak saliency label is inaccurate. Each sample is presented with three columns: the input pedestrian image on the left, the activation from the primary ReID head in the middle, and the weak label for the auxiliary saliency detection head on the right. The gradient descent directions for the two objectives are contradictory.

To address this fundamental limitation, we propose to break through the dilemma by calibrating the conflicting gradient yield by the auxiliary objective with that from the primary objective as a reference. Specifically, When $\boldsymbol{g_a}$ is conflicting with $\boldsymbol{g_p}$, we consider $\boldsymbol{g_p}$ as a reference and manually alter the direction of $\boldsymbol{g_a}$ by mapping it to the normal plane of $\boldsymbol{g_p}$ to get the calibrated gradient $\boldsymbol{g_a^c}$ as

$$\boldsymbol{g_a^c} = \boldsymbol{g_a} - \frac{\boldsymbol{g_a} \cdot \boldsymbol{g_p}}{\|\boldsymbol{g_p}\|^2} \boldsymbol{g_a}, \qquad \text{subject to } (\boldsymbol{g_a} \cdot \boldsymbol{g_p}) < 0, \quad (6)$$

Remark: This procedure changes the direction of the conflicting gradient to ensure it does not conflict with the primary task. With the calibrated gradient, the model can consider the partial guidance of the auxiliary objective, ensuring the joint effort is non-conflicting with the primary objective. It is effective in minimizing the side effects caused by the inaccurate labeling of the auxiliary task while still performing standard first-order gradient descent to optimize the model.

3.4. Deployment-Time Optimization

We further formulate the PAOA+ to exploit the data characteristic of the target domain and perform deployment time optimization with the available samples during testing. Considering that the proposed PAOA is composed of a shared feature encoder f_θ and two separate task heads h_p and h_a that are optimized jointly during model training. When the trained model is deployed in a new environment, given a batch of identity-unknown samples $\{x'_i\}_{i\in\{1,\cdots,B'\}}$, with the corresponding weakly labels $\{y_i'^{(a)}\}$ generated by the pre-trained saliency detection model, the shared feature extract f_θ can be further optimized on the auxiliary task by minimizing the following loss

$$\mathcal{L}_{\text{test}} = \frac{1}{B} \sum_{1}^{B} \mathcal{L}_{\text{aux}}(x'_i, y_i'^{(a)}; f_\theta). \quad (7)$$

So that f_θ can be swiftly adapted by considering the data distribution of the new environment, further to yield improved performance on the main task. Note the difference from domain adaptation based methods which assume the test sample is available during the training phase for explicit distribution alignment, PAOA+ only requires a batch of samples with arbitrary numbers for on-the-fly updates, allowing it to seamlessly adapt to new data distributions.

3.5. Model Training and Deployment

Training stage: Given the formulation of the primary and auxiliary tasks, the PAOA model is designed in multitask learning architecture and can benefit from the conventional learning supervision by jointly minimizing the primary and auxiliary losses. The parameters are iteratively optimized with the training loss (Eq. (5)). As the feature extractor parameterized by θ is shared by both the primary and auxiliary tasks, it will be jointly updated with two gradients: $\boldsymbol{g_p}$ for the primary task and $\boldsymbol{g_a}$ for the auxiliary task. To seek positive interactions between tasks, the direction of $\boldsymbol{g_a}$ will be calibrated only if it conflicts with $\boldsymbol{g_p}$ by Eq. (6). Note that the cross-entropy loss provides stronger supervision for person classification, therefore we use its gradients as the reference to calibrate that of the auxiliary task. This calibrated gradient ensures the auxiliary task is harmoniously trained with the primary task by back-propagation and thereby brings benefits to facilitate the deployment-time optimization. The overall training procedure is depicted in Algorithm 1.

Deployment stage: To make a consistent comparison with DG ReID methods, we can directly apply the trained PAOA model for identity representation extraction. Additionally, the improved PAOA+ model further performs deployment time optimization during the testing stage to mitigate the domain shift between the training and testing domains. Given the identity representations, subsequent identity retrieval is performed by a general distance metric.

4. Experiment

4.1. Experimental Settings

(a) CUHK-SYSU (b) CUHK03 (c) MSMT17 (d) Market1501

Figure 4. Example identity samples from different domains and its corresponding weak labels for the auxiliary task. Significant domain gaps are caused by the variation on nationality, illumination, viewpoints, resolution, scenario, etc. As complementary, the pedestrian saliency label can provide a guide on the most discriminative person area.

Algorithm 1 Model Training with PAOA regularization

Input: Labeled dataset $\mathcal{D} = \{(x_i, y_i^{(p)})\}$ for primary task, weak label generator $\mathcal{G}$ for auxiliary task, shared feature extractor f_θ, head modules h_p/h_a for primary/auxiliary tasks.
Output: Trained f_θ, h_p and h_a.
for $i = 1$ **to** max_iter **do**
Randomly sample a mini-batch $\{(x_i, y_i^{(p)})\}_{i\in\{1,\cdots,N_\text{B}\}}$ from source dataset $\mathcal{D}$.
Generate the weak label for the auxiliary task by $\{y_i^{(a)} = \mathcal{G}(x_i)\}_{i\in\{1,\cdots,N_\text{B}\}}$.
Compute the training loss (Eq. (5)) and calculate the gradients.
Calibrate the conflicting gradients (Eq. (6)).
Update the network by gradient descent.
end for

Implementation Details We used PFAN [41] as the wake label generator for the auxiliary task. The shared feature extractor is a ResNet50 [10] pre-trained on ImageNet [6] to bootstrap the feature discrimination. The balancing hyper-parameter in Eq. (5) was set to 0.1. The batch size was set to 64, including 4 images for 16 randomly sampled identities. All images were resized to 128×256. The model was trained for 200 epochs with the Adam optimizer [17]. The learning rate was set to $3.5e-4$. The dimension of the extracted identity representation was set to 2048. The dimension of the saliency map is 64×32. The learning rate for PAOA+ was set to $1e-6$ and the test batch size was 200. The post-optimization step is set to 1 for balancing performance and efficiency. All the experiments were implemented on PyTorch [27] on a single A100 GPU.

Datasets and Evaluation Protocol We conducted multi-source domain generalized ReID on a wide range of benchmarks. including Market1501 (M) [43], MSMT17 (MS) [34], CUHK03 (C3) [20], CUHK-SYSU (CS) [35], CUHK02 (C2) [19], VIPeR [8], PRID [12], GRID [24], and iLIDs [44]. We evaluated the performance of PAOA on the four small-scale datasets following the traditional setting [2, 15, 30, 38]. We also performed leave-one-out evaluations by using three datasets for training and the remaining for the test [4, 22, 42]. Note that the CUHK-SYSU is only for training given all the images are captured by the same camera. To learn a discriminative model benefits from diverse identities, all the identities regardless of the original train/test splits, were used for training. We adopted Mean average precision (mAP) and Rank-1 of CMC as the evaluation metrics.

4.2. Comparison with SOTA methods

We compared the proposed PAOA against several recent SOTA methods, and the comparison results are shown in Table 1 and Table 2. Under a fair comparison with existing DG ReID methods, the PAOA model outperforms all the competing methods by a significant margin on both the traditional setting and the large-scale settings across all the evaluation metrics. It shows a clear advantage over the recent SOTA methods. Notably, even trained with fewer datasets compared with [2, 16, 30], the proposed method is still able to extract discriminative features for identity matching. Besides, we extended our analysis to include the results from the test-time optimization variant, PAOA+, which notably improves PAOA consistently across all benchmarks. These results provide additional evidence on the effectiveness of the associative learning strategy, where the auxiliary task can promote the primary ReID objective during test time given the absence of identity labels.

4.3. Ablation Studies

Component Analysis We investigated the effects of different components in PAOA model design to study their individual contributions. The baseline model is a ResNet50 pre-trained on ImageNet. The comparison results are shown in Table 3, from which we can observe that the auxiliary objective and the gradient calibration strategies can consistently improve performance. With further deployment-time optimization, our model can be advanced by benefiting from mining the data characteristics in the target domain. It is notable that the variant without gradient calibration can always

Table 1. Comparison with the SOTA methods on traditional evaluation protocol. The best results are shown in **red** and the second-best results are shown in blue.

Source	Method	PRID		GRID		VIPeR		iLIDs		Average	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
M+D+C2 +C3+CS	DIMN [30]	52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2	57.9	47.5
	SNR [16]	66.5	52.1	47.7	40.2	61.3	52.9	89.9	84.1	66.3	57.3
	DMG-Net [2]	68.4	60.6	56.6	51.0	60.4	53.9	83.9	79.3	67.3	61.2
M+C2+ C3+CS	M3L [42]	64.3	53.1	55.0	44.4	66.2	57.5	81.5	74.0	66.8	57.2
	MetaBIN [4]	70.8	61.2	57.9	50.2	64.3	55.9	82.7	74.7	68.9	60.5
	ACL [38]	73.5	63.0	65.7	55.2	75.1	66.4	86.5	81.8	75.2	66.6
	META [36]	71.7	61.9	60.1	52.4	68.4	61.5	83.5	79.2	70.9	63.8
	PAOA (Ours)	74.0	65.6	67.2	56.3	76.6	66.7	87.1	83.1	76.2	67.9
	PAOA+ (Ours)	**75.1**	**66.5**	**67.8**	**56.9**	**77.2**	**67.7**	**88.0**	**83.9**	**77.0**	**68.8**

Table 2. Comparison with the SOTA methods on large-scale evaluation protocol. The best results are shown in **red** and the second-best results are shown in blue.

Method	Reference	M+MS+CS→C3		M+CS+C3→MS		MS+CS+C3→M		Average	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
SNR [16]	CVPR2020	17.5	17.1	7.7	22.0	52.4	77.8	25.9	39.0
QAConv$_{50}$ [22]	ECCV2020	32.9	33.3	17.6	46.6	66.5	85.0	39.0	55.0
M^{3}L [42]	CVPR2021	35.7	36.5	17.4	38.6	62.4	82.7	38.5	52.6
MetaBIN [4]	CVPR2021	43.0	43.1	18.8	41.2	67.2	84.5	43.0	56.3
ACL [38]	ECCV2022	49.4	50.1	21.7	47.3	76.8	90.6	49.3	62.7
META [36]	ECCV2022	47.1	46.2	24.4	52.1	76.5	90.5	49.3	62.9
PAOA	Ours	49.8	50.5	25.1	51.5	77.1	90.8	50.7	64.3
PAOA+	Ours	**50.3**	**50.9**	**26.0**	**52.8**	**77.9**	**91.4**	**51.4**	**65.0**

Table 3. Effects on mAP (%) value of the proposed modules. Aux: auxiliary objective. GC: gradient calibration. DTO: deployment-time optimization.

Aux	GC	DTO	C3	MS	M	Average
✗	✗	✗	42.8	20.5	73.1	45.5
✓	✗	✗	44.8	20.9	73.5	46.4
✓	✗	✓	47.0	23.1	75.2	48.4
✓	✓	✗	49.8	25.1	77.1	50.7
✓	✓	✓	**50.3**	**26.0**	**77.9**	**51.4**

benefit more from that post-optimization compared with the PAOA+ model, This further illustrates that the referenced calibration mechanism has already enabled the PAOA model to be more attentive to the domain-invariant pedestrian region, and therefore it relies less on on-the-fly optimization.

Table 4. Effects on mAP (%s) of update iterations during deployment optimization.

Dataset	0	1	2	3	4
C3	49.8	50.3	50.5	50.6	50.3
MS	25.1	26.0	26.5	26.0	25.0
M	77.1	77.9	77.5	77.0	76.2
Avg.	50.7	51.4	51.5	51.2	50.5

Gradient Calibration Designs We adopted a primary-referenced design for the gradient calibration between the primary and auxiliary objectives. This was based on the fact that the primary instance classification objective provides stronger supervision to identify pedestrians, while the auxiliary objective is to guide the instance classifier to attentively focus on the pedestrian area and ignore the domain-specific interference. It's weakly labeled and therefore is intricately noisy which can lead to a negative influence on the primary objective, reflected by the conflicting gradient. We examined the effect of the calibration design by additionally testing three more formulations as demonstrated in Figure 5. Table 5 shows the auxiliary-referenced design yielded the worst performance, given the gradients of the auxiliary objective is noisy and unreliable, using it as the reference is harmful to the learning of the primary objective. By contrast, the mutually referenced calibration design includes the primary gradients as referenced on top of the auxiliary-referenced design, which alleviates the fallout caused by the gradient destruction, despite it's still inferior to the baseline. In comparison, the primary-referenced design consistently obtained improved performance which supports the design of the proposed primary referenced gradient calibration.

Update iterations for deployment-time optimization We analyzed the influence of update iterations for optimizing the

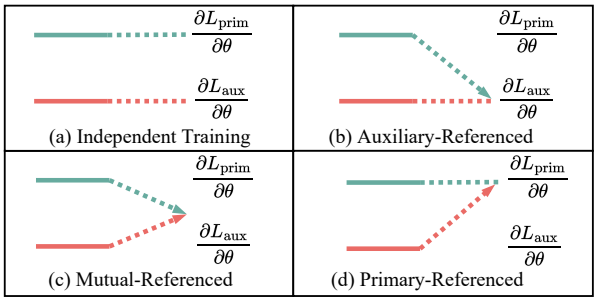


Figure 5. Illustration of different gradient calibration designs. (a) No gradient calibration as [29]. (b) Gradients of the primary objective are calibrated with the auxiliary objective as a reference. (c) Gradients are calibrated in relation to each other as a reference, as designed in [37]. (d) Gradients of the auxiliary objective are calibrated with the primary objective as a reference.

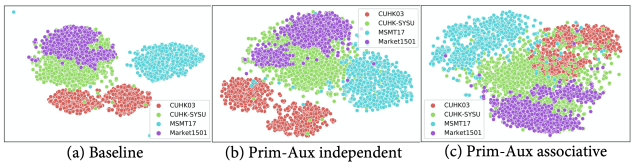
Table 5. Comparison of different gradient calibration designs by mAP (%). Refer to Figure 5 for the corresponding design.

Design	C3	MS	M	Avg.
a	44.8	20.9	73.5	46.4
b	44.1	21.7	74.7	46.8
c	47.3	23.0	75.3	48.5
d	**49.8**	**25.1**	**77.1**	**50.7**

model with all test samples in deployment time. Ablating with iterations from 0 to 4 (Table 4), we noted consistent performance improvement by updating the model at the initial steps. This is attributed to the auxiliary objective guiding swift adaptation to the test domain. This improvement is attributed to the auxiliary objective facilitating rapid adaptation to the test domain. However, excessive updates result in a model forgetting issue by overwhelming the extractor with the auxiliary. Notably, deployment-time optimization is more effective for target datasets (*i.e.*, MSMT17) with larger domain shifts, which further proves that target-aware updates that mitigate domain shifts more effectively. Balancing efficiency and effectiveness, PAOA+ adopts single-step updates across all datasets to attain the global optimal solution.

Visualization We visualized the pedestrian images and the model activation maps to intuitively illustrate the effectiveness of PAOA. We took the feature map of the final convolutional layer ($4th$ layer) as the activation map, and compared the baseline model with the proposed PAOA. As can be observed in Figure 6, the PAOA model can accurately be attentive to the pedestrian area, while the baseline model is partially focus and some discriminative areas are missed. This is benefited from the auxiliary objective, as shown in the second column, which provides assistive supervision on instance classification learning. Therefore, PAOA model can extract discriminative yet generic identity representation for ReID. To also visualized the TSNE distribution of the extracted feature representations in Figure 7. The target domain is Market1501 and the model was trained with other three source domains. Training independently with the auxiliary objective can condense the feature space compared with the baseline, however it's still prone to domain shift, especially for CUHK03. As a comparison, the proposed PAOA can significantly reduce domain shifts with a much more compact feature space.

Figure 6. Visualization of activation maps. For each pedestrian image, the four columns from left to right are: (1) Person image, (2) Weak label for auxiliary objective, (3) Activation map from the proposed PAOA model, (4) Activation map from the baseline. The proposed PAOA helps the model be more attentive on the pedestrian region to learning domain-invariant representation.



(a) Baseline (b) Prim-Aux independent (c) Prim-Aux associative

Figure 7. TSNE visualization on extracted features. 200 samples were randomly sampled from each domain. Learning with jointly the primary and auxiliary objectives can condense the feature distribution. The proposed model which associates the primary and auxiliary objectives can derive a more compact feature space.

5. Conclusions

In this work, we introduced a novel *Primary-Auxiliary Objectives Association* (PAOA) regularization to learn a generalizable ReID model for extracting domain-unbiased representations more generalizable to unseen novel domains for person ReID. PAOA encourages the model to get rid of the interference of domain-specific knowledge and to learn from discriminative pedestrian information by the association of learning an auxiliary pedestrian detection objective with a primary instance classification objective. To mitigate the fallout caused by the noisy auxiliary labels, we further derive a referenced-gradient calibration strategy to alter the gradient of the auxiliary object when it's conflicting with the primary object. The PAOA framework is task-agnostic, making it readily adaptable to other tasks through the incorporation of a close auxiliary task and a shared learning module.

Acknowledgements

This work was supported by the China Scholarship Council, the Alan Turing Institute Turing Fellowship, Veritone. This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT.

References

[1] Eugene PW Ang, Lin Shan, and Alex C Kot. Dex: Domain embedding expansion for generalized person re-identification. In *BMVC*, 2021. 2

[2] Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, and Ling-Yu Duan. Person30k: A dual-meta generalization network for person re-identification. In *CVPR*, 2021. 6, 7

[3] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational visual media*, 5(2), 2019. 3

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

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

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[7] Cian Eastwood, Ian Mason, Christopher KI Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. In *ICLR*, 2022. 3

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

[9] Ke Han, Chenyang Si, Yan Huang, Liang Wang, and Tieniu Tan. Generalizable person re-identification via self-supervised batch norm test-time adaption. In *AAAI*, 2022. 3

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

[11] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. In *ECCV*, 2020. 2, 3

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

[13] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *NeurIps*, 2021. 3

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

[15] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature alignment and restoration for domain generalization and adaptation. *arXiv*, 2020. 6

[16] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, 2020. 2, 6, 7

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[18] Qilei Li, Jiabo Huang, and Shaogang Gong. Local-global associative frame assemble in video re-id. *BMVC*, 2021. 1

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

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

[21] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 1

[22] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *ECCV*, 2020. 6, 7

[23] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *NeuIPS*, 2021. 3

[24] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *IJCV*, 2010. 6

[25] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *ECCV*, 2020. 1

[26] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, 2021. 1

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[28] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 2

[29] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 2, 3, 8

[30] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, 2019. 6, 7

[31] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 3, 4

[32] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A Olshausen, and Trevor Darrell. Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 3

[33] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022. 3

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

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

[36] Boqiang Xu, Jian Liang, Lingxiao He, and Zhenan Sun. Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 372–388. Springer, 2022. 7

[37] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *NeurIps*, 2020. 3, 5, 8

[38] Pengyi Zhang, Huanzhang Dou, Yunlong Yu, and Xi Li. Adaptive cross-domain learning for generalizable person re-identification. In *ECCV*, 2022. 6, 7

[39] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 3

[40] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, 2020. 1

[41] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019. 4, 6

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

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

[44] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, 2009. 6

[45] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 1

[46] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv*, 2021. 1

[47] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020. 1, 2

[48] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, 2020. 2

[49] Zijie Zhuang, Longhui Wei, Lingxi Xie, Haizhou Ai, and Qi Tian. Camera-based batch normalization: an effective distribution alignment method for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1), 2021. 2

[50] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *ECCV*, 2020. 2