Dynamically Transformed Instance Normalization Network for Generalizable Person Re-Identification

Bingliang Jiao[1,2,3,5], Lingqiao Liu[4], Liying Gao[1,2,3], Guosheng Lin[5†], Lu Yang[1,2,3], Shizhou Zhang[1,3], Peng Wang[1,2,3†], and Yanning Zhang[1,3†]

[1] School of Computer Science, Northwestern Polytechnical University, China
[2] Ningbo Institute, Northwestern Polytechnical University, China
[3] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean, China
[4] The University of Adelaide, Australia
[5] Nanyang Technological University, Singapore

Abstract. Existing person re-identification methods often suffer significant performance degradation on unseen domains, which fuels interest in domain generalizable person re-identification (DG-PReID). As an effective technology to alleviate domain variance, the Instance Normalization (IN) has been widely employed in many existing works. However, IN also suffers from the limitation of eliminating discriminative patterns that might be useful for a particular domain or instance. In this work, we propose a new normalization scheme called Dynamically Transformed Instance Normalization (DTIN) to alleviate the drawback of IN. Our idea is to employ dynamic convolution to allow the unnormalized feature to control the transformation of the normalized features into new representations. In this way, we can ensure the network has sufficient flexibility to strike the right balance between eliminating irrelevant domain-specific features and adapting to individual domains or instances. We further utilize a multi-task learning strategy to train the model, ensuring it can adaptively produce discriminative feature representations for an arbitrary domain. Our results show a great domain generalization capability and achieve state-of-the-art performance on three mainstream DG-PReID settings.

Keywords: Person Re-identification, Domain Generalization, Instance Normalization, Dynamic Convolution.

1 Introduction

Person re-identification (PReID) aims at matching identical persons across different cameras. Many supervised PReID methods have recently achieved promising success when training and evaluating images under the same environment. However, the performance of these methods tends to significantly degrade when

†*Corresponding author*,
bingliang.jiao@mail.nwpu.edu.cn, *peng.wang@nwpu.edu.cn*, *ynzhang@nwpu.edu.cn*

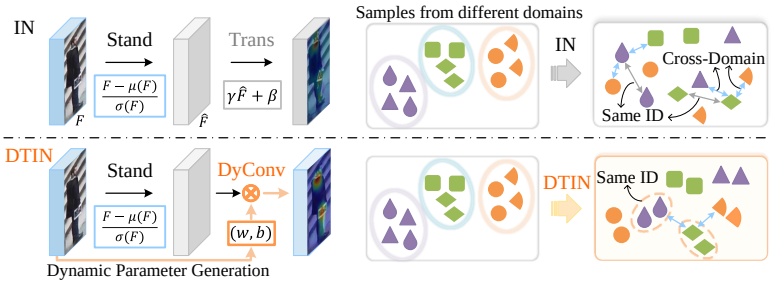

Fig. 1. The sketches of the traditional Instance Normalization (IN) and our proposed Dynamically Transformed Instance Normalization (DTIN). Stand, Trans, and DyConv represent feature standardization, affine transformation, and dynamic convolution operation, respectively. Generally, IN can alleviate domain variances between inputs by removing their statistical contrast but inevitably eliminates discriminative information of inputs. It thus leads to a relatively large distance between features of the same pedestrian (in the upper case). Therefore, in our DTIN, we employ unnormalized features to guide the transformation of normalized features into new representations to adapt individual domains and instances. In this way, features extracted by our DTIN can both generalize well and be distinguishable.

testing images from an unseen environment. It is a common belief that the change of capturing environment, e.g., change of illumination, view-angles, and seasons causes the domain shift, and existing approaches are not robust under those changes. For this reason, domain generalizable person re-identification (DG-PReID), which aims to build a ReID model that could be more robust to the domain shift and work in an unseen environment, has received increasing attention.

Recently a series of DG-PReID methods [5, 10, 4, 12, 16] have been proposed. Among these prior works, Instance Normalization (IN) [23] is widely used to produce domain-invariant features by removing the statistical contrast across feature channels. However, the removed statistics not only encode irrelevant domain-specific patterns but also contain discriminative patterns that may be useful for performing ReID for a particular domain or instance. To address this issue, Pan *et al.* [19] propose an IBN-Net which concatenates features extracted by IN and Batch Normalization (BN) together. A more sophisticated method SNR [12] tries to identify the useful information discarded by IN and then compensate it back. In this work, we aim to address the same issue as the aforementioned works but address the problem with a different principle. Rather than focusing on compensation, we shift our attention to building a module that is sufficiently flexible for learning a mapping function to combine both normalized features and unnormalized features. Our insight is that when the network is sufficiently flexible, we could use end-to-end training to discover a model that can strike a balance between eliminating irrelevant domain-specific features and adapting to individual domains or instances.

To this end, in this work, we propose a new normalization scheme named Dynamically Transformed Instance Normalization (DTIN). The main idea of our DTIN is to employ unnormalized features to guide the transformation of normalized features into a new representation that is adapted to the current domain and instance. To do so, as shown in Figure 1, we integrate IN with a dynamic instance-aware convolution operation (DyConv). More specifically, in our DTIN, the adaptive parameters of DyConv are generated under the guidance of unnormalized features and then used to re-calibrate and transform the normalized ones. The intuition is that the unnormalized features can transform the normalized ones to dynamically capture information useful to distinguish instances in specific domains. In this way, we could achieve good generalization by adaptively creating feature representation for each individual domain or instance. In addition to utilizing unnormalized features as the control signal for the transformation, we further design a dynamic control path that makes it convenient for the network to utilize patterns from multiple layers. This design further adds the flexibility of learning a generalizable mapping function. To train the network, we adopt a multi-task learning formulation to encourage the network to generate representations that work well for an arbitrary training domain.

2 Related works

Domain Adaptation for Person Re-Identification Unsupervised domain adaptation (UDA) requires that the deep models trained on the source domain can adapt to the target domain and work well on it. Generally, in the UDA setting, numbers of unlabeled target domain data are allowed to be accessed during the training phase. Recently, generation-based methods [17, 22] and fine-tuning methods [26, 29] become two major solutions for UDA. The former type of method mainly employs style transfer algorithms like CycleGAN [34] to transfer the style of labeled source domain data to the style of the target domain. By training with these transferred data, ReID models can be adapted to target domains. In addition, fine-tuning methods try to allocate pseudo labels for unlabelled target domain data by clustering. In this way, they can fine-tune ReID models on the target domain with obtained pseudo labels. Although the UDA methods have the potential to adapt ReID models to a target domain, they highly rely on the unlabeled target domain data, which are not always available in real-world applications.

Generalizable Person Re-Identification. Domain Generalizable (DG) Person Re-Identification (PReID) aims to train a robust and generalizable person re-identification model which can perform well on unseen target domains without a further update. Recently, many relevant works [5, 4, 12, 16, 10] have been proposed to achieve this goal. Among these methods, Instance Normalization (IN) [23] has been widely employed to alleviate domain variances between input features by removing their statistical contrast. For instance, Jia *et al.* [10] simply inserts IN into the early layers of the backbone model to eliminate domain disparity. However, IN also inevitably causes discriminative information loss for

input features, which limits its application. Therefore, many revisions have been given based on IN. For instance, an SNR module [12] has been proposed to disentangle identity-relevant information from features discarded by IN and reintroduce it back. In addition, Choi *et al.* [4] propose a Meta Batch-Instance Normalization (MetaBIN) model which integrates IN with Batch Normalization and balances their effort with a set of learnable trade-off parameters. However, these methods are often based on the principle of disentangling the domain-relevant information and domain-invariant information to design the model structure, which is perhaps a more challenging problem than domain generalization. Unlike the prior works, we explore dynamically transforming normalized features into appropriate representations to make them adaptive to individual domains and instances.

3 Proposed Methods

This section elaborates on our proposed Dynamically Transformed Instance Normalization (DTIN) module and multi-task training strategy. We first provide the preliminaries of Instance Normalization and Dynamic Convolution, which underpins our proposed method.

3.1 Preliminaries

Let $F \in \mathbb{R}^{C\times H\times W}$ be the convolutional feature of a given input image I, where C is the channel dimension and H, W represent the height and width of the feature map, respectively.

Instance Normalization (IN) is firstly proposed for the style transformation task [23] to remove style information from input features. Recently, it has been widely used for the DG-PReID task to extract domain-invariant representations by removing their statistical contrast across feature channels. Generally, IN consists of a standardization component and an affine transformation component, which can be respectively written as,

$$\begin{aligned} \text{Standardization: } \widehat{F} &= \frac{F-\mu(F)}{\sigma(F)} \\ \text{Affine transformation: } \widetilde{F} &= \gamma\widehat{F}+\beta, \end{aligned} \tag{1}$$

where $\widetilde{F}$ represents the normalized features, μ and σ denote channel-wise mean and standard deviation of F; γ and β are trainable affine parameters learned from end-to-end training on the entire training dataset. By removing domain-specific factors encoded in the mean and standard deviation of the feature channels, IN can effectively enhance the robustness of ReID models by making them less sensitive to domain change.

The effectiveness of IN in extracting domain-invariant representations has been verified in many existing domain generalization studies [10, 19, 4]. However, the classical IN approach also faces a significant drawback. As also discussed in [19], the channel-wise variance contrast removed by IN also encodes

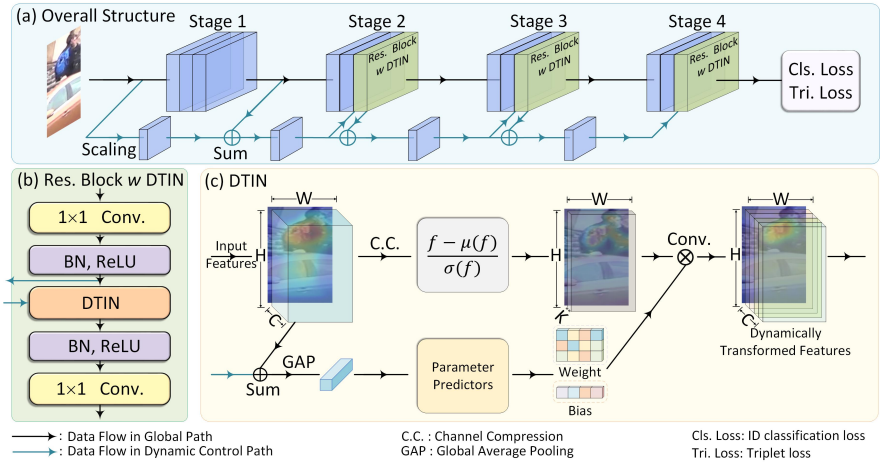

Fig. 2. The sketch of our Dynamically Transformed Instance Normalization (DTIN). The scaling modules in (a) are a set of 3×3 convolutional layers, which are responsible for matching the spatial and channel scales between multi-level features. The parameter predictors are used to generate adaptive parameters, *i.e.*, weight and bias, for the dynamic convolution operation.

discriminative patterns that might be useful for performing ReID in a particular domain or instance. Blindly removing them will have a negative impact on the DG-PReID task.

Dynamic Convolution (DyConv) can adaptively adjust the feature extraction paradigm according to the input instances. Compared to static modules, DyConv is more easily adapted to out-of-distribution inputs due to its flexiblility [31, 1]. The idea of DyConv is firstly proposed in [11]. In a dynamic convolutional layer, the weight and bias of each filter are generated from the input and applied to the input. Formally, it can be written as:

$$F^{DC} = \text{conv}\left(F; w(F)\right), \tag{2}$$

where F^{DC} represents the output of DyConv; $\text{conv}(\cdot)$ indicates the convolution operation, $w(F)$ represents the filter parameters, *i.e.*, weight and bias, which are generated from input features F. So, $w(F)$ is a function of F. DyConv is an ideal structure for model adaptation since it can adjust the model parameters from the input. Motivated by that, this paper employs DyConv to modulate the normalized feature after IN.

3.2 Dynamically Transformed Instance Normalization

In this work, we propose to integrate DyConv into IN to overcome the drawback of IN. The intuition of doing so is to adaptively re-calibrate normalized features

via DyConv into new representations that can adapt to individual domains and instances. To this end, in this work, we propose a Dynamically Transformed Instance Normalization (DTIN) module, as illustrated in Figure 2 (c). In our DTIN, we employ a 1×1 dynamic convolution [11] to transform normalized features extracted from IN.

Intuitively, we can regard the filters generated in a DyConv as pattern detectors [3] that can adaptively adjust their sensitivity towards different visual patterns based on input image content. Also, in our design, we modify DyConv by using the unnormalized features to generate filters to process normalized (IN) features. In this way, we could make most of both normalized and unnormalized features for identifying the useful features for ReID. Please see Figure 4 for some concrete examples that show the advantage of the proposed design. More formally, the DTIN operation can be written as:

$$\begin{aligned} F^{DTIN} &= \text{conv}(\frac{F^c - \mu(F^c)}{\sigma(F^c)}; w(F')), \\ F^c &= \text{conv}(F; \theta_0),\ w(F') = \text{fc}(\text{relu}(\text{fc}(\text{pool}((F')), \theta_1)), \theta_2), \end{aligned} \tag{3}$$

where the F and the F^{DTIN} are input and output features of our DTIN; $\text{conv}(F; \theta_0)$ indicates a 1×1 convolutional layer which is employed in our DTIN to reduce the channel dimension of features before IN from C to K (K=64); $\text{pool}(\cdot)$ denotes spatially average pooling operation; θ_1 and θ_2 represent parameters of fully connected layers $\text{fc}(\cdot)$. $w(F')$ denotes the produced adaptive parameters (including weight and bias); F' is the features employed for parameter prediction, which involves the unnormalized features and will be explained in the following with more details;

Note that there is no need to apply an additional static affine transformation to standardized features $\frac{F^c - \mu(F^c)}{\sigma(F^c)}$ as in IN since DyConv has already performed a dynamic transformation to them.

Details of the control signal F'. As mentioned above, we modify DyConv by generating the filters from F' rather than F. In our design, F' contains features before the IN layer, i.e., unnormalized features. This is motivated by our concern that IN will eliminate some domain-specific patterns that might be useful for performing ReID for an individual domain or instance. For this reason, our design is different from an architecture of using DyConv layer after IN. In the latter case, the filters are not generated from unnormalized features but from normalized features. We compare this alternative in Figure 3 and find that it leads to much worse performance.

In addition to using unnormalized features for producing F', we also use skip-connections to bring signals from low-level features to enrich F'. We call this design the dynamic control path, and the visualization of this scheme can be found in Figure 2 (a). As seen, the dynamic control path aggregates features from multiple levels, and a set of 3×3 convolutional layers are employed along the dynamic control path to ensure the consistency of feature map dimensions. After integrating these multi-level features to produce F', we fed it into DyConv to generate w as in Equation. 3.

Advantages of our DTIN. Our DTIN integrates the IN and DyConv modules. IN plays a role in eliminating irrelevant domain-specific patterns but is with the limitation of removing discriminative domain-specific or instance-specific information. DyConv module controlled by the unnormalized features can not only re-calibrate those features to avoid over-normalization but also transform the normalized features (through convolution operation) into an appropriate representation. In this sense, our DTIN allows a flexible mapping function that could achieve a trade-off between eliminating ineffective domain-specific factors and adapting to individual domains or instances.

The Deployment of DTIN. To fairly evaluate the effectiveness of our designed DTIN, in this work, we insert it into the widely employed ResNet-50 model [8]. The empowered model is named as **DTIN-Net**, as shown in Figure 2 (a). The ResNet-50 consists of four major stages, and each stage contains different numbers of residual blocks. To avoid introducing excess computational consumption, we only use our DTIN to replace the 3×3 convolution layer in the last residual block of the 2nd - 4th stages, to refine features extracted in these stages. Besides, we experimentally insert vanilla IN after the first residual stage to facilitate convergence [16].

3.3 Multi-task Training Strategy

Generally, we can access data from multiple source domains at the DG-PReID training stage. Here we denote all the available source domains as $\mathbf{D} = \{D_s\}_{s=1}^{S}$, where we use the S to indicate the number of available domains. In addition, $D_s = (x_i^s, y_i^s)_{i=1}^{N_s}$ represents the s-th domain, the x_i^s and y_i^s respectively represent the input image and the label of i-th sample, and the N_s denotes the number of instances in the s-th domain.

Our aim is to learn a feature extractor to adaptively generate feature representations that can achieve good ReID performance for arbitrary domain. Thus, we treat each domain as an independent task and supply those tasks with a shared feature extractor. This is equivalent to a multi-task-style training process.

Formally, for each domain, we create an ID classifier φ_s to perform ID classification. The s is the domain index. The label space of s-th classifier φ_s is the identities in the s-th domain. We also apply triplet loss to the samples randomly sampled from the s-th domain. The overall objective function of our multi-task training strategy is to minimize the average of loss in each domain:

$$L_{M.T.} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N_s}\Big(\sum_{i=1}^{N_s} L_{ce}\left(\varphi_s\left(\psi(x_i^s)\right), y_i^s\right) + L_{tri.}(\psi(x_i^s), \psi(x_{p^i}^s), \psi(x_{n^i}^s))\Big), \tag{4}$$

where the i and s are indexes of image and domain; N_s represents the number of instances in the s-th domain; $x_{p^i}^s$ and $x_{n^i}^s$ are sampled positive instance and negative instance for x_i^s; $\psi(\cdot)$ denotes the feature extraction model, *i.e.*, our DTIN-Net; the L_{ce} and $L_{tri.}$ represent cross-entropy loss and triplet loss.

Note that in contrast to the aforementioned multi-task training, one could also stack the training samples from all domains together. Then it is possible to apply ID classification loss with a classifier with the label space corresponding to all identities and also apply the triplet loss to triplets sampled from all samples. However, we empirically find this scheme leads to worse performance. The reason is that by doing so, we have the risk of encouraging the network to use features that distinguish domains but not instances. For example, if the negative image pairs in the triplet loss are from two different domains, the network could partially rely on such features to pull those negative image pairs apart.

4 Experiments

4.1 Implementation Details and Evaluation Setting

Dataset. In this paper, we employ the mainstream person ReID datasets to evaluate the generalization capability of our method, including 6 larger datasets Market1501 (M) [27], DukeMTMC-ReID (D) [30], Cuhk02 (C2) [13], Cuhk03 (C3) [14], CuhkSYSU (CS) [25], MSMT17 (MS) [24], and 4 smaller ones termed VIPeR (V) [7], PRID (P) [9], GRID (G) [18], and QMUL i-LIDS (Q) [28].

Settings. In this paper, we first adopt two widely used multiple source domain generalization protocols as in [5] to evaluate the generalization capability of our model. In Protocol-1, we employ the M, D, C2, C3, and CS to construct the training set and evaluate our model on the V, P, G, and Q, respectively. In Protocol-2, the training set comprises M, D, C3, and MS, and the test set is the same as Protocol-1. Besides, we also follow a cross-domain setting [4], in which we train our model on M (D) and test it on D (M).

Implementation Details. Before the domain generalization training, we firstly pre-train our DTIN-Net on the ImageNet [6] dataset. Under all protocols mentioned above, we train our model for 120 epochs. Particularly, for the cross-domain setting, we do not use the multi-task training strategy since it only contains one source domain. Besides, in our DTIN-Net, the dynamic control path receives and aggregates features before the first residual stage, features after the first residual stage, and features before our DTIN modules. The learning rate is initialized as 3.5×10^{-4} and divided by 10 at the 40-th and 70-th epochs, respectively. In all experiments, each image is resized to 256×128 for training and test. During the training phase, each image is flipped horizontally with a probability of 0.5. All results reported in this section are the mean of two repetitive experiments. In addition, random erasing is employed for data augmentation. The widely used metrics CMC and mAP are employed to evaluate our model.

4.2 Comparison with State-of-the-art Methods

To clarify, in the remainder of this paper, the "Baseline" model indicates a ResNet-50 model trained with ID classification loss and triplet loss.

Multiple Source Domain Generalization. We compare our DTIN-Net with other state-of-the-art methods under the aforementioned Protocol-1 and

Table 1. The re-identification performance comparison between our DTIN-Net and other DG-PeReID algorithms under the Protocol-1 (P-1) and Protocol-2 (P-2) settings. It can be found that our DTIN-Net outperforms the compared algorithms under both settings.

Setting	Method	GRID		VIPeR		PRID		i-LIDS		Average	
		mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1	mAP	CMC-1
P-1	DIMN [21]	41.1	29.3	60.1	51.2	52.0	39.2	78.4	70.2	57.9	47.5
	DMG-Net [2]	56.6	*51.0*	60.4	53.9	68.4	60.6	83.9	79.3	67.3	61.2
	RaMoE [5]	54.2	46.8	64.6	56.6	67.3	57.7	**90.2**	**85.0**	69.1	61.5
	MetaBIN [4]	*57.9*	48.4	*68.6*	*59.9*	**81.0**	**74.2**	87.0	81.3	*73.6*	*66.0*
	DTIN-Net	**60.6**	**51.8**	**70.7**	**62.9**	*79.7*	*71.0*	*87.2*	*81.8*	**74.6**	**66.9**
P-2	SNR [12]	41.3	30.4	65.0	55.1	60.0	49.0	*91.9*	*87.0*	64.6	55.4
	DMG-Net [2]	47.2	37.3	70.9	62.3	*69.7*	*59.7*	88.2	83.0	69.0	60.6
	RaMoE [5]	*53.9*	*43.4*	**72.2**	*63.4*	66.8	56.9	**92.3**	**88.4**	*71.3*	*63.0*
	DTIN-Net	**58.4**	**49.4**	*71.9*	**64.0**	**77.4**	**67.8**	89.2	85.3	**74.2**	**66.6**

Table 2. The performance of single domain generalization for person re-identification. Compared with other state-of-the-art algorithms, our DTIN-Net achieves a promising performance, which indicates our algorithm is also effective when training with limited data.

Method	M → D		D→ M	
	mAP	CMC-1	mAP	CMC-1
IBN-Net [19]	24.3	43.7	23.5	50.7
OSNet [33]	25.9	44.7	24.0	52.2
CrossGrad [20]	27.1	48.5	26.3	56.7
QAConv [15]	28.7	48.8	27.2	58.6
L2A-OT [33]	29.2	50.1	30.2	63.8
OSNet-AIN [32]	30.5	52.4	30.6	61.0
SNR [12]	33.6	55.1	33.9	66.7
MetaBIN [4]	33.1	55.2	35.9	69.2
DTIN-Net	**36.1**	**57.0**	**37.4**	**69.8**

Protocol-2 settings, and the results are shown in Table 1. As we can find, on almost all datasets, our DTIN-Net model achieves comparable or better performance than compared methods. Only on the i-LIDs dataset our DTIN-Net is worse than RaMoE [5]. Nevertheless, our DTIN-Net model outperforms all other compared algorithms in the average performance, which demonstrates the superiority of our proposed method.

Cross-Domain Generalization. To further evaluate the generalization capability of our DTIN-Net, we additionally compare it with other state-of-the-art algorithms under the cross-domain generalization setting. The experiential results are shown in Table 2. Here, we do not compare our DTIN-Net with other multi-source DG-PReID methods [5, 2] since these methods design their structure and training strategy based on the premise that more than one source domain can be accessed. Compared to the best competitors, *i.e.*, SNR [12] and MetaBIN [4], our DTIN-Net achieves significantly better performance, 2.8%

Table 3. The effectiveness of our designed components, namely dynamically transformed instance normalization module (DTIN) and multi-task training strategy (M.T.). It can be found that our design is effective in enhancing the generalization capability of ReID models.

	DTIN	M.T.	GRID		VIPeR	
			mAP	CMC-1	mAP	CMC-1
Baseline	×	×	46.5	37.2	65.3	56.4
Base+IN	×	×	52.7	42.9	67.5	58.7
DTIN-Net	✓	×	$56.4^{\uparrow \mathbf{3.7}}$	$46.2^{\uparrow \mathbf{3.3}}$	$69.1^{\uparrow \mathbf{1.6}}$	$60.4^{\uparrow \mathbf{1.7}}$
DTIN-Net	✓	✓	$\mathbf{60.6}^{\uparrow \mathbf{7.9}}$	$\mathbf{51.8}^{\uparrow \mathbf{8.9}}$	$\mathbf{70.7}^{\uparrow \mathbf{3.2}}$	$\mathbf{62.9}^{\uparrow \mathbf{4.2}}$

mAP on average (M → D). The superiority of our DTIN-Net could come from two aspects. Firstly, inheriting the advantages of dynamic convolution operation [11] and instance normalization [23], our DTIN module naturally adapts to out-of-distribution inputs and is robust to domain variance. Secondly, our DTIN integrates these two effective modules in a judicious manner. By enhancing the flexibility of ReID model, our DTIN-Net can adaptively balance normalizing domain interference and adapting to individual domains or instances. In this way, our DTIN can extract features that generalize well on unseen domains.

4.3 Ablation Studies

All experiments in this subsection are based on the protocol-1 setting. To clarify, in this subsection, "Base+IN" indicates the model (based on ResNet-50) inserted with classical IN layers before the convolutional layers we replace our DTIN with. "Base+IN+DyConv" is the two-module setup that simply concatenates IN with a 1×1 dynamic convolution module [11].

Table 4. The comparison between versions of DTIN-Net with or without our designed dynamic control path (DyCtrl). M.T. represents the multi-task training strategy. As we can find, the DyCtrl consistently enhances the generalization capability of our model.

	M.T.	DyCtrl	GRID		VIPeR	
			mAP	CMC-1	mAP	CMC-1
DTIN-Net	×	×	54.7	44.1	68.6	59.6
DTIN-Net	×	✓	$56.4^{\uparrow \mathbf{1.7}}$	$46.2^{\uparrow \mathbf{2.1}}$	$69.1^{\uparrow \mathbf{0.5}}$	$60.4^{\uparrow \mathbf{0.8}}$
DTIN-Net	✓	×	58.8	49.3	70.0	62.3
DTIN-Net	✓	✓	$60.6^{\uparrow \mathbf{1.8}}$	$51.8^{\uparrow \mathbf{2.5}}$	$70.7^{\uparrow \mathbf{0.7}}$	$62.9^{\uparrow \mathbf{0.6}}$

Effectiveness of Designed Components. To evaluate the effectiveness of our designed DTIN and multi-task training strategy (M.T.), we gradually add them to the "Baseline" model and compare the performance. For a fair comparison, in Table 3, we also give the performance of the "Base+IN" model. It can be found that the robustness of the "Baseline" model to domain variance

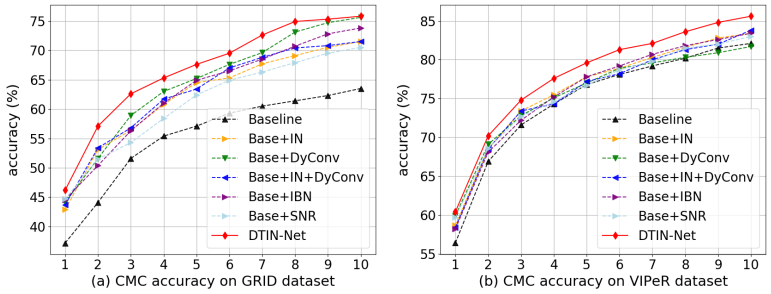

Fig. 3. The CMC-1 to CMC-10 accuracy comparison of variations on the DTIN-Net architecture. The ("Base+DyConv", "Base+SNR", and "Base+IBN") are the versions replacing our DTIN module with a 1×1 Dynamic Filter module [11], SNR module [12], and IBN module [19]. The "Base+IN+DyConv" is the two-module setup which simply concatenates IN with a 1×1 dynamic convolution module [11]. For a fair comparison, all these models above are trained without our multi-task training strategy.

can be relatively improved by simply applying IN to it. However, the inherent drawback of IN, *i.e.*, the loss of discriminative information, limits the capability of the "Base+IN" model. For our DTIN-Net, it achieves a significantly better recognition accuracy than the "Base+IN" model on unseen domains, about 3.7% mAP and 1.6% mAP on GRID and VIPeR datasets, respectively. The interpretation could be that the delicate parameter prediction strategy employed by our DTIN provides sufficient semantic information, based on which our DTIN can adaptively calibrate normalized features to adapt to individual domains and instances. Besides, the generalization capability of our DTIN-Net can be further improved (4.2% mAP on the GRID dataset) if we additionally utilize M.T. to train our DTIN-Net. The interpretation of the improvement could be that our training strategy can guide our DTIN to learn appropriate transformations to adapt normalized features to arbitrary domains.

Effectiveness of Dynamic Control Path. To ensure that the features to predict adaptive parameters contain sufficient semantic information, in our DTIN-Net, a dynamic control path (DyCtrl) is given to integrating low-level features as the semantic supplement to high-level features. As shown in Table 4, this straightforward operation brings 2.5% and 0.6% CMC-1 improvement to our DTIN-Net on GRID and VIPeR dataset (trained with M.T.), respectively. It indicates that additional semantic information provided by our designed DyCtrl indeed benefits to construct instance-adaptive calibration for each input and improves its distinguishability on unseen domains.

Comparison between Our DTIN and Other Relevant Modules. In Figure 3, we give the performance of models inserting traditional IN ("Base+IN"), or dynamic convolution module [11] ("Base+DyConv") at the positions we set our DTIN to. As we can find, all these single-module setups improve the

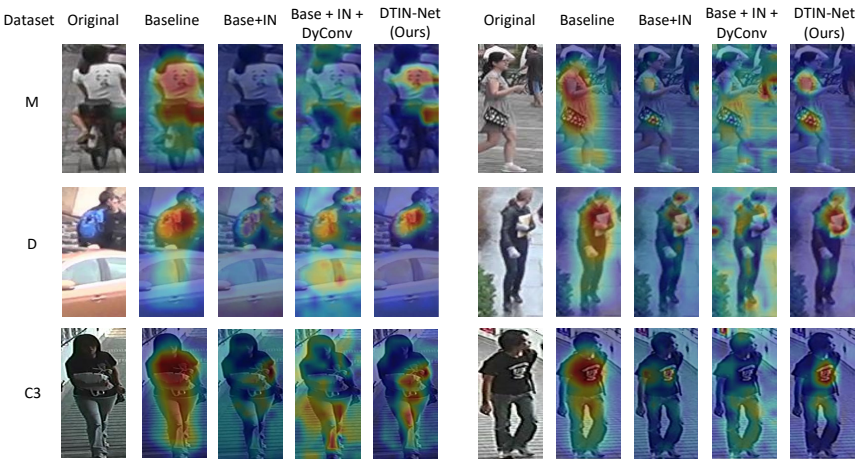

Fig. 4. The activation maps of features extracted by "Baseline" model, "Base+IN" model, two-module setup "Base+IN+DyConv" and our DTIN-Net. The M, D, C3 indicate cases in those rows are sampled from Market1501, DukeMTMC-ReID, Cuhk03 datasets. It can be found that the IN modifies the original activation map. While correctly reducing the response values for irrelevant regions, such as the "car windows" region, it also lowers the contrast between discriminative regions and background regions. Directly combining IN with DyConv, i.e., "Base+IN+DyConv", fails to overcome the limitation of IN and seems to highlight the background incorrectly. In contrast, our DTIN can correctly locate the identity-relevant patterns and remove the activations from the irrelevant regions.

"Baseline" model. However, simply concatenating these two effective modules ("Base+IN+DyConv") together does not bring an additional improvement. The interpretation could be that the information loss caused by IN may seriously limit the flexible nature of the dynamic module. As shown in Figure 4, the "Base+IN+DyConv" even seems to highlight the background incorrectly due to the insufficient semantic perception capability of its generated parameters. On the contrary, thanks to the delicate dynamic control path, our DTIN ensures features to parameter prediction contain sufficient semantic information. In this way, our DTIN can effectively calibrate normalized features to adapt to individual domains and instances. As shown in Figure 4, our DTIN can effectively capture discriminative clues of each instance (like the cartoon pattern in the first case in Market1501 dataset) and thus is able to re-calibrate normalized features to be distinguishable. In addition, in Figure 3 we also give the performance of models replacing our DTIN with SNR [12] ("Base+SNR") and IBN [19] ("Base+IBN") module. Thanks to the flexible nature of our DTIN, it can adaptively strike the right balance between normalizing irrelevant domain features and adapting to individual domains or instances and thus achieves a better performance than these compensation-based methods.

Table 5. The comparison of model capacity and computation consumption between our DTIN-Net and other state-of-the-art methods. Our DTIN-Net achieves a significantly better generalization performance than compared methods with comparable parameters (Params) and floating-point operations (FLOPs).

Method	GRID		VIPeR		Params	FLOPs
	mAP	CMC-1	mAP	CMC-1	(M)	(GMac)
Base+IN	52.3	42.5	67.3	58.4	**23.5**	4.1
RaMoE [5]	54.2	46.8	64.6	56.6	39.3	4.1
MetaBIN [4]	57.9	48.4	68.6	59.9	23.6	4.1
DTIN-Net	**60.6**	**51.8**	**70.7**	**62.9**	25.5	**3.7**

Analysis about Model Capacity. In this work, we employ a dynamic convolution module to instance-adaptively transform normalized features to make them adapt to individual domains and instances. However, it also raises the suspicion, namely, whether the superiority of our DTIN-Net is caused by increasing the capacity of the backbone model with the computation-expensive dynamic convolution modules? To explore this suspicion, we compare the generalization capability and model capacity of our DTIN-Net with other state-of-the-art models. In addition, the capacity of the "Base+IN" model is also given for comparison. The results are given in Table 5, from which we can summarize two important findings. Firstly, thanks to the light-weight design of our DTIN-Net, our DTIN-Net achieves a significantly better performance than the "Base+IN" model, about 8.3 % mAP on GRID dataset, by increasing only a few additional parameters (2 M) and even saving 9.8 % calculation (0.4 GMac). Secondly, compared with other state-of-the-art methods, *i.e.*, RaMoE and MetaBIN, our DTIN-Net achieves a better generalization performance (averagely, 4.6% mAP on GRID dataset) with comparable parameters and computations. It indicates that the superiority of our DTIN-Net on DG-PReID task is not caused by improving model capacity.

T-SNE Visualization Results. To intuitively show how our DTIN enhances the generalization capability of ReID models, in Figure 5 we exhibit the t-SNE visualization of features extracted by the "Baseline" model, the "Base+IN" model, and our DTIN-Net. Here, we randomly sample 60 pairs of query and gallery images from each of the 4 unseen domains (GRID, PRID, VIPeR, and i-LIDS). In this experiment, the perplexity and iteration of t-SNE is set to 6 and 50. respectively. The features used for visualization are extracted after the last DTIN module for our model and the corresponding position for comparing models. As shown in Figure 5 (a), we can find that the "Baseline" model is seriously influenced by domain factors, which causes clear domain boundaries. Particularly, in the selected gray section, since samples from query set and gallery set are captured under different cameras, there even exists a significant gap between the features of query samples and gallery samples. For the "Base+IN" in Figure 5 (b), it effectively erases the domain interference of each input and thus breaks the domain boundaries. However, due to the loss of discriminative information,

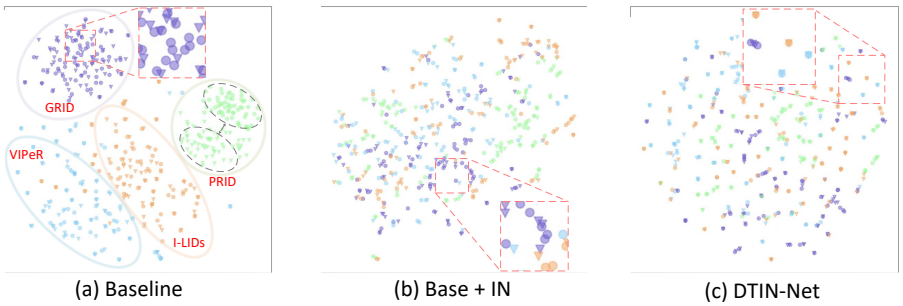

Fig. 5. The t-SNE results of features on four unseen target datasets (VIPeR, PRID, GRID, and i-LIDS). For comparison, we simultaneously give the visualization of features processed by the "Baseline" model, the "Base+IN" model, and our DTIN-Net. Dots with identical colors are from the same domain. Besides, the triangular and circular dots represent query and gallery, respectively.

large intra-class distances exist between query instances and gallery instances (as shown in the selected section) which may limit the recognition accuracy of ReID models. For our DTIN module, by applying instance-level adaptive calibration for each normalized feature, it significantly improves the distinguishability of normalized features by IN. Specifically, we can find that our DTIN-Net forms a relatively independent sub-cluster for each person instance with clear boundaries, while the cluster boundaries in "Base-IN" are blurred.

5 Conclusion

In this work, we propose a new normalization scheme named Dynamically Transformed Instance Normalization (DTIN), which effectively alleviates the inherent drawback of Instance Normalization that inevitably impacts discriminative information of input features. By transforming normalized features into appropriate representation in a learnable and adaptive manner, our DTIN empowers ReID models to strike the right balance between normalizing irrelevant domain features and adapting to individual domains or instances. In addition, we further propose a multi-task formulation on multiple training domains to train our model. Extensive experiments demonstrate the superiority of our proposed method.

Acknowledgments This work is supported by National Key R&D Program of China (No.2020AAA0106900), the National Natural Science Foundation of China (No.U19B2037), Shaanxi Provincial Key R&D Program (No.2021KWZ-03), Natural Science Basic Research Program of Shaanxi (No.2021JCW-03). This work is also supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2018-003), and the MOE AcRF Tier-1 research grant: RG95/20.

References

1. Akula, A., Jampani, V., Changpinyo, S., Zhu, S.C.: Robust visual reasoning via language guided neural module networks. Proc. Advances in Neural Inf. Process. Syst. (2021)
2. Bai, Y., Jiao, J., Ce, W., Liu, J., Lou, Y., Feng, X., Duan, L.Y.: Person30k: A dual-meta generalization network for person re-identification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2123–2132 (2021)
3. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5659–5667 (2017)
4. Choi, S., Kim, T., Jeong, M., Park, H., Kim, C.: Meta batch-instance normalization for generalizable person re-identification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 3425–3435 (2021)
5. Dai, Y., Li, X., Liu, J., Tong, Z., Duan, L.Y.: Generalizable person re-identification with relevance-aware mixture of experts. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 16145–16154 (2021)
6. Deng, J., Dong, W., Socher, R., Li, L., Li, K., FeiFei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 248–255 (2009)
7. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proc. Eur. Conf. Comp. Vis. pp. 262–275 (2008)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 770–778 (2016)
9. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on Image analysis. pp. 91–102 (2011)
10. Jia, J., Ruan, Q., Hospedales, T.M.: Frustratingly easy person re-identification: Generalizing person re-id in practice. arXiv preprint arXiv:1905.03422 (2019)
11. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. Proc. Advances in Neural Inf. Process. Syst. **29**, 667–675 (2016)
12. Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 3143–3152 (2020)
13. Li, W., Wang, X.: Locally aligned feature transforms across views. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 3594–3601 (2013)
14. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014)
15. Liao, S., Shao, L.: Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In: Proc. Eur. Conf. Comp. Vis. pp. 456–474 (2020)
16. Lin, S., Li, C.T., Kot, A.C.: Multi-domain adversarial feature generalization for person re-identification. IEEE Trans. Image Process. **30**, 1596–1607 (2020)
17. Liu, J., Zha, Z.J., Chen, D., Hong, R., Wang, M.: Adaptive transfer network for cross-domain person re-identification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 7202–7211 (2019)
18. Loy, C.C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 1988–1995. IEEE (2009)

19. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via IBN-Net. In: Proc. Eur. Conf. Comp. Vis. pp. 464–479 (2018)
20. Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S.: Generalizing across domains via cross-gradient training. arXiv preprint arXiv:1804.10745 (2018)
21. Song, J., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)
22. Tang, Y., Yang, X., Wang, N., Song, B., Gao, X.: CGAN-TM: A novel domain-to-domain transferring method for person re-identification. IEEE Trans. Image Process. **29**, 5641–5651 (2020)
23. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
24. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 79–88 (2018)
25. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. arXiv preprint arXiv:1604.01850 (2016)
26. Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., Tian, Y.: AD-Cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2020)
27. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 1116–1124 (2015)
28. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: Proc. British Machine Vis. Conf. pp. 1–11. No. 6 (2009)
29. Zheng, Y., Tang, S., Teng, G., Ge, Y., Liu, K., Qin, J., Qi, D., Chen, D.: Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 8371–8381 (2021)
30. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*. In: Proceedings of the IEEE international conference on computer vision. pp. 3754–3762 (2017)
31. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization in vision: A survey. arXiv preprint arXiv:2103.02503 (2021)
32. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. (2021)
33. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Learning to generate novel domains for domain generalization. In: Proc. Eur. Conf. Comp. Vis. pp. 561–578 (2020)
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. IEEE Int. Conf. Comp. Vis. (2017)