Part-Aware Transformer for Generalizable Person Re-identification

Hao Ni[1] Yuke Li[1] Lianli Gao[2] Heng Tao Shen[1] Jingkuan Song[2*]

[1] University of Electronic Science and Technology of China (UESTC)
[2]Shenzhen Institute for Advanced Study, UESTC
{haoni0812, liyuke65535, jingkuan.song}@gmail.com,

Abstract

Domain generalization person re-identification (DG-ReID) aims to train a model on source domains and generalize well on unseen domains. Vision Transformer usually yields better generalization ability than common CNN networks under distribution shifts. However, Transformer-based ReID models inevitably over-fit to domain-specific biases due to the supervised learning strategy on the source domain. We observe that while the global images of different IDs should have different features, their similar local parts (e.g., black backpack) are not bounded by this constraint. Motivated by this, we propose a pure Transformer model (termed Part-aware Transformer) for DG-ReID by designing a proxy task, named Cross-ID Similarity Learning (CSL), to mine local visual information shared by different IDs. This proxy task allows the model to learn generic features because it only cares about the visual similarity of the parts regardless of the ID labels, thus alleviating the side effect of domain-specific biases. Based on the local similarity obtained in CSL, a Part-guided Self-Distillation (PSD) is proposed to further improve the generalization of global features. Our method achieves state-of-the-art performance under most DG ReID settings. The code is available at `https://github.com/liyuke65535/Part-Aware-Transformer`.

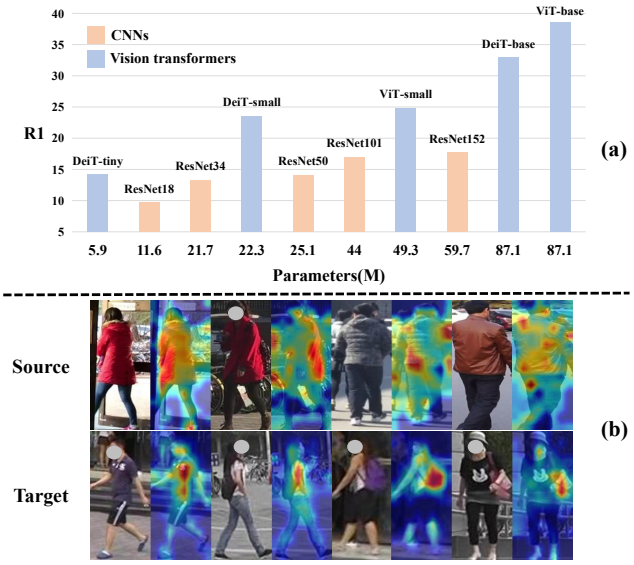

Figure 1. (a) We applied different Transformers to DG ReID. Models are trained on Market and tested on MSMT. Results show Vision transformers (blue bars) are better than CNNs (orange bars) even with fewer parameters. (b) Visualization of attention maps of "class token" on source domain (MSMT) and target domain (Market). We use ViT [4] as the backbone and fuse the attention results of the shallow layers. However, the attention to discriminative information is still limited on target domain.

1. Introduction

Person Re-Identification (ReID) [35, 2, 38, 37] aims to find persons with the same identity from multiple disjoint cameras. Thanks to the great success of Convolutional Neural Network (CNN) in the field of computer vision [11, 24], supervised, unsupervised person ReID has made significant progress. However, a more challenging task, domain generalization (DG) ReID [31] which trains a model on source domains yet generalizes well on unseen target domains, still lags far behind the performance of the supervised ReID.

Thus, many DG methods are proposed to learn generic features. These methods explore the generalization of CNN based on disentanglement [10, 19] or meta-learning [3, 18]. Recently, Transformer has gained increasing attention in computer vision. It is a neural network based on attention mechanisms [26]. Vision Transformer usually yields better generalization ability than common CNN networks under distribution shift[32, 17]. However, existing pure transformer-based ReID models are only used in supervised and pre-trained ReID [16, 8]. The generalization of Transformer is still unknown in DG ReID.

To investigate the performance of Transformer in DG

*Jingkuan Song is the corresponding author.

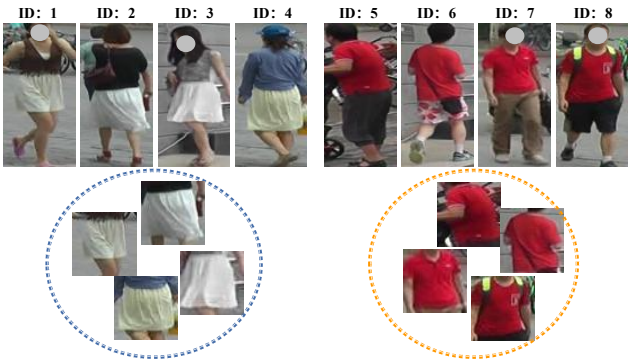

Figure 2. Local similarity among parts with different IDs. It comes from the visual data themselves, not from ID labels.

ReID, we use different Transformers and CNNs as backbones to test their cross-domain performance from Market to MSMT. The results show that Vision Transformers are much better than common CNNs, as shown in Figure 1 (a). Even with fewer parameters, Transformers still outperform CNNs. For instance, DeiT-tiny [25] with 5.9M parameters is much better than ResNet18 with 21.7M parameters. Despite the great performance of ViT [4], we still experimentally find that the attention to discriminative information is limited to the unseen target domain. As shown in Figure 1(b), some discriminative information on the target domain is ignored, such as the black backpack and grey coat.

The above phenomenon shows that Transformer- based ReID models inevitably overfit to domain-specific biases due to the supervised learning strategy on the source domain. It is manifested in the insufficient learning of local information on unseen target domains. We observe that while the global images of different IDs should have different features, their similar local parts (e.g., White skirt, red T-shirt) are not bounded by this constraint, as shown in Figure 2. And these ID-independent local similarities can provide extra visual knowledge from the images themselves. Ignoring this similarity leads the Transformer to focus on the ReID task instead of learning generic features, resulting in more over-fitting to domain-specific biases.

To this end, we design a proxy task, named Cross-ID Similarity Learning (CSL), to mine local similarities shared by different IDs and learn generic features without using ID labels. CSL is based on part-aware attention to learn discriminative information across different IDs. The part-aware attention concatenates the "part token" and the "image tokens" in the region of interest to learn local representations. In each mini-batch, we use a memory bank to calculate the distance between the current local features and the samples of the entire dataset to mine apparent local similarity. The apparent similarity is learned not from ID annotations, but from the visual data themselves [30]. Thus it allows the model to learn generic features because it only cares about the visual similarity of the parts regardless of the ID labels, thus alleviating the side effect of domain-specific biases.

Part-guided Self-Distillation (PSD) is proposed to further improve the generalization of the global representation. Self-distillation has been proven effective in DG image classification [27, 33]. It can learn visual similarities beyond hard labels and make the model converges easier to the flat minima. However, we experimentally find that the traditional self-distillation method would reduce the generalization in DG ReID. The reason is that ReID is a fine-grained retrieval task, and the difference between different categories is not significant. It is difficult to mine useful information from the classification results. Therefore, PSD uses the results of CSL to construct soft labels for global representation. In general, the motivation of self-distillation is similar to CSL, which is to learn generic features by data themselves regardless of the ID labels.

Extensive experiments have proved that CSL and PSD can improve the generalization of the model. Specifically, our method achieves state-of-the-art performance under most DG ReID settings, especially when using small source datasets. Under the Market→CUHK-NP setting, our method exceeds state-of-the-art by 3.2% and 4.6% in Rank1 and mAP, respectively. The contributions of this work are three-fold:

a) We propose a pure Transformer-based framework for DG ReID for the first time. Specifically, we design a proxy task, named Cross-ID Similarity Learning module (CSL), to learn generic features.

b) We design part-guided self-distillation (PSD) for DG ReID, which learns visual similarities beyond hard labels to further improve the generalization.

c) Extensive experiments have proved that our Part-aware Transformer achieves state-of-the-art of DG ReID.

2. Related Work

2.1. Domain Generalizable Person ReID.

Supervised and unsupervised domain adaptation person ReID have achieved great success. But DG ReID is still a challenging task. It requires the model to train a model on source domains yet generalize to unseen target domains. Due to its huge practical value, it has been widely studied in recent years. The concept of DG ReID was first proposed in [31]. [22, 9] applied meta-learning to learn domain-invariant features. [10] proposed to disentangle identity-irrelevant information. Last but not the least, [21] proposed IBN-net to explore the effect of combining instance and batch normalization, which was widely used in later DG ReID methods due to its good transferability and effectiveness. However, pure Transformer does use batch normalization, so IBN cannot bring gain to our model. But even

without using IBN, our method still outperforms the existing CNN-based state-of-the-art in DG ReID.

2.2. Transformer-related Person ReID.

The original Transformer is proposed in [26] for natural language processing (NLP) tasks. Based on ViT [4], [8] applies pure Transformer to supervised ReID for the first time, which introduces side information to improve the robustness of features. [16] further proposed self-supervised pre-training for Transformer-based person ReID, which mitigates the gap between the pre-training and ReID datasets from the perspective of data and model structure.

Recently, some work investigate the generalization of vision Transformers [32]. In DG ReID, TransMatcher [14] employs hard attention to cross-matching similarity computing, which is more efficient for image matching. However, it still uses CNN as the main feature extractor, and the role of the Transformer is mainly reflected in image matching. Our method is the first to investigate the generalization ability of pure Transformer in DG ReID.

2.3. Proxy Task and Self-Distillation.

Proxy Task and Self-Distillation have been extensively studied, and we only discuss their contribution to generalization here.

Proxy Task is referred to as learning with free labels generated from the data itself, such as solving Jigsaw puzzles [20], predicting rotations [6] or reconstruction [5]. Since these tasks are not related to the target task (such as image classification), they can guide the model to learn generic features, which leads to less over-fitting to domain-specific biases [1]. CSL picks similar parts from the entire dataset without using ID labels, thus encouraging the model to learn the discriminative information shared by different IDs.

Self-Distillation (SD) uses soft labels containing "richer dark knowledge", which can reduce the difficulty of learning the mapping and further improve the generalization ability of the model [27]. Besides, [33] found that SD can help models converge to flat minima, improving the generalization of features. However, traditional SD methods are not suitable for ReID. Because it is a fine-grained retrieval task. So we propose PSD to replace traditional methods.

3. Methodology

We proposed a pure Transformer-based framework, named Part-aware Transformer (PAT), to learn generalizable features, as shown in Figure 3. In the following, we describe the main components of our method. First, we introduce our Transformer encoder composed of L blocks, which simultaneously extracts global and local features (Sec. 3.1). Next, we design a proxy task, named Cross-ID Similarity Learning (CSL), to learn the generic features (3.2). It mines local similarity shared by different IDs and encourages model to learn generic features, thereby reducing over-fitting on source datasets. Finally, a Part-guided Self-Distillation module (PSD) is proposed to further improve the generalization of global features (Sec. 3.3). It constructs soft labels based on the similarity of local features, which solves the problems existing in traditional self-distillation methods on ReID. CSL and PSD are jointly trained in an end-to-end manner (Sec. 3.4).

3.1. Transformer Encoder

Our Transformer encoder f consists of L blocks. Each block contains global attention, part-aware attention and a feed-forward network. Global/Part-aware attention is used to extract global/local features.

Input of Transformer Encoder. We split an input image $x \in \mathbb{R}^{H\times W\times C}$ into non-overlapping N patches by a patch embedding module. Each patch is treated as an "image token" $\{x_i|i=1,2,\ldots,N\}$. Besides, a learnable "class token" x_{cls} and three "part tokens"$\{x_{p_i}|i=1,2,3\}$ are concatenated with all "image tokens". Then, the input to the Transformer encoder can be expressed as:

$$\mathcal{Z} = [x_{cls}, x_{p_1}, x_{p_2}, x_{p_3}, x_1, \ldots, x_N] + \mathcal{P} \tag{1}$$

where $\mathcal{Z}$ represents input sequence embeddings, $\mathcal{P} \in \mathbb{R}^{(N+4)\times D}$ is position embedding. D is the number of channels.

The attention mechanism is based on a trainable associative memory with query Q, key K, and value V. They are all computed from the vector sequence $\mathcal{Z}$, which can be formulated as:

$$\begin{aligned} Q &= \mathcal{Z}W_Q = [q_{cls}, q_{p_1}, q_{p_2}, q_{p_3}, q_1, \ldots, q_N] \\ K &= \mathcal{Z}W_K = [k_{cls}, k_{p_1}, k_{p_2}, k_{p_3}, k_1, \ldots, k_N] \\ V &= \mathcal{Z}W_V = [v_{cls}, v_{p_1}, v_{p_2}, v_{p_3}, v_1, \ldots, v_N] \end{aligned} \tag{2}$$

where W_Q, W_K, W_V are different linear transformations.

Global Attention. To extract global features, we use "class token" and all "image tokens" to perform global attention. The output matrix of global attention can be obtained by:

$$Attention(Q_{cls},K_{cls},V_{cls}) = Softmax(\frac{Q_{cls}K_{cls}^T}{\sqrt{D}})V_{cls} \tag{3}$$

where $Q_{cls} = [q_{cls}, q_1, \ldots, q_N]$, $K_{cls} = [k_{cls}, k_1, \ldots, k_N]$ and $V_{cls} = [q_{cls}, q_1, \ldots, q_N]$. Then the outputs of global attention are sent to FFN network. Repeat this process L times and we get a global feature $f(x_{cls})$.

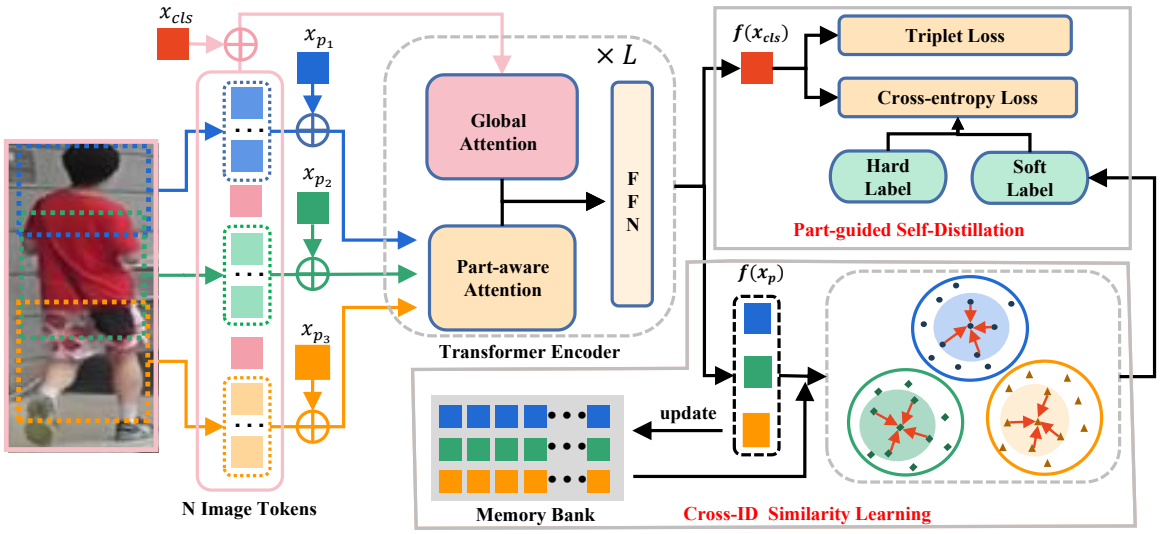

Figure 3. Illustration of our proposed method. FFN is a feed-forward network. We split an input image into N non-overlapping patches as "image tokens". The input of the model includes a "class token" (x_{cls}), three "part tokens" ($\{x_{p_i}|i \in 1,2,3\}$) and N "image tokens". The "class token" concatenates all "image tokens" to get a global feature through global attention. For each "part token", it concatenates "image tokens" of the region of interest to obtain the local feature belonging to this region. The output of the L Transformer block includes a global feature $f(x_{cls})$ and three local features $\{f(x_{p_i})|i \in 1,2,3\}$. We use three local representations and samples in the memory bank to solve a proxy task, named Cross-ID Similarity Learning (CSL). The results of the CSL guide the global representation to perform Part-guided Self-Distillation (PSD).

Part-aware Attention. To learn local similarity from data themselves, we need to extract local features using part-aware attention. For each "part token" x_{p_i}, we use x_{p_i} and "image tokens" belonging to a special region to perform part-aware attention. the output matrix of part-ware attention can be obtained by:

$$Attention(Q_{p_i}, K_{p_i}, V_{p_i}) = Softmax(\frac{Q_{p_i} K_{p_i}^T}{\sqrt{D}}) V_{p_i} \quad (4)$$

where $Q_{p_i} = [q_{p_i}, q_{k_i}, \ldots, q_{k_i+m}]$ and $\{k_i, \ldots, k_i + m\}$ represent the serial number of $(m+1)$ "image tokens" associated with the part token x_{p_i}. Q_{p_i} and K_{p_i} are handled in the same way. That is, "part tokens" will only interact with "image tokens" in the region of interest. In this work, We take three overlapping square areas along the vertical direction, as shown in the figure 3. The outputs of part-ware attention are sent to the FFN network. Repeating this process L times we get three local features $\{f(x_{p_i})|i = 1,2,3\}$.

3.2. Cross-ID Similarity Learning

Existing Transformer-based ReID models perform representation learning based on global attention with annotated images, which leads the models to focus too much on domain-specific information, thus models get over-fitted on source domains. Besides, transformer-based ReID models ignore local similarities between different IDs, which can be helpful for generic representation learning. Specifically, cross-ID local similarities are ID-irrelevant, they offer visual knowledge to the models regardless of labels. Just like self-supervised learning, it provides the model with additional visual knowledge from the images themselves, not their labels.

To learn generic features, we propose a proxy task named Cross-ID Similarity Learning (CSL). Our method is based on the observation that although the entire images of different IDs are quite different, the local regions of some IDs are similar, such as red short sleeves, white skirts, and many more (see Figure 2). The above information is not enough to discriminate ID on the source domain, but it is helpful to learn generic features. Just like other self-supervised learning methods in DG, such as solving jigsaw puzzles and predicting rotations, solving proxy tasks allow our model to learn generic features regardless of ReID task, and hence less over-fitting to domain-specific biases.

Since it is difficult to find local similarities shared by different IDs in a mini-batch, we need to compare as many samples as possible in one gradient descent. To this end, we maintain a momentum-updated memory bank $\{w_{p_i}|i = 1,2,3\}$ for each part token x_{p_i} during learning, which can be expressed as:

$$w_{p_i}^j = \begin{cases} f\left(x_{p_i}^j\right) & t = 0, \\ (1-m) \times w_{p_i}^j + m \times f\left(x_{p_i}^j\right) & t > 0 \end{cases} \quad (5)$$

where t is the training epoch, m is the momentum and $j =$

$1, \ldots, K$. K is the number of samples in the source dataset.

For each local feature $f(x_{p_i}^j)$ in current mini-batch, we compare it with the entire memory bank. We select k local features closest to $f(x_{p_i}^j)$ from $\{w_{p_i}\}^K$ to form a set $\{\mathcal{K}_{p_i}^j\}_{j=1}^k$ of positive samples. Then, the distance between positive samples and $f(x_{p_i}^j)$ is minimized by softmax-clustering loss , which can be formulated as:

$$\mathcal{L}_{p_i}^j = -log\frac{\sum_{w_{p_i}^m \in \{\mathcal{K}_{p_i}^j\}_{j=1}^k} \exp(\frac{f(x_{p_i}^j)w_{p_i}^m}{\tau})}{\sum_{n=1}^K \exp(\frac{f(x_{p_i}^j)w_{p_i}^n}{\tau})} \quad (6)$$

where τ is a temperature coefficient. Minimizing $\mathcal{L}_{p_i}$ encourages the model to pull similar patches $\{\mathcal{K}_{p_i}^j\}_{j=1}^k$ close to $f(x_{p_i}^j)$ while pushing dissimilar patches away from $f(x_{p_i}^j)$ in feature space. In this way, the model can learn those visually similar patches in different IDs and make the Transformer notice the regions where this useful information is located.

3.3. Part-guided Self-Distillation

Self-distillation has been shown to help improve generalization. For example, it can learn visual similarities beyond hard labels and make the model converges easier to the flat minima [27, 33, 23]. However, our experiments found that the traditional self-distillation method could not improve the generalization of ReID. Because traditional self-distillation method relies on the output of the classifier to get the similarity between different categories. This is effective in image classification because of the large visual differences between different classes. For example, cats and dogs are visually similar, but cats and tables are very dissimilar. Such information can be utilized to facilitate learning. However, ReID is a fine-grained retrieval task, and the differences between different IDs are insignificant. So there is no useful information in the classification results.

To this end, we propose Part-guided Self-Distillation (PSD), which uses the visual similarity of local parts to implement self-distillation. In Section 3.2, each local representation $f(x_{p_i}^j)$ gets k positive samples, and an image includes three part. Therefore, there are $3k$ positive samples in total for the global representation. We regard the IDs $\{I_i\}_{i=1}^{3k}$ corresponding to these $3k$ part as similar IDs. Soft labels Y_s^j of $f(x_{cls}^j)$ are constructed as follows:

$$Y_s^j|_i = \begin{cases} 1-\alpha & i = y_s^j \\ \frac{\alpha}{3k} n_i & i \in \{I_i\}_{i=1}^{3k}, \\ 0 & i \notin \{I_i\}_{i=1}^{3k} \cup \{y_s^j\} \end{cases} \quad (7)$$

where α is the weight of similar categories, n_i is the number of the i-th ID in $\{I_i\}_{i=1}^{3k}$ and y_s^j is ground truth of $f(x_{cls}^j)$. That is, the more similar parts, the greater the probability of the ID to which these parts belong.

Then, the part-guided self-distillation loss can be formulated as:

$$\mathcal{L}_s^j = -\lambda Y_s^j \log P\Big(f(x_{cls}^j)\Big) - (1-\lambda) Y^j \log P\Big(f(x_{cls}^j)\Big) \quad (8)$$

where Y^j is the one-hot hard label, λ is the coefficient to balance soft label and P is the classifier that predicts probability distribution on source dataset. Since the apparent similarity is obtained by comparing the local representations of the current sample with the entire dataset. Therefore, it reflects the similarity between IDs better than the classification result of the classifier.

3.4. Loss Function and Discussion

Loss function. In addition to the softmax-clustering loss and part-guided self-distillation loss used in 3.2 and 3.3, we also use the triplet loss with soft-margin to constrain the distance between positive and negative sample pairs. it can be formulated as follows:

$$\mathcal{L}_{tri} = log[1 + exp(||f_a - f_p||_2^2 - ||f_a - f_n||_2^2)] \quad (9)$$

where $\{f_a, f_p, f_n\}$ are the features of a triplet set.

For a mini-batch with N_b samples, our entire loss function can be expressed as:

$$\mathcal{L} = \mathcal{L}_{tri} + \sum_{j=1}^{N_b}(\sum_{i=1}^{M} \mathcal{L}_{p_i}^j + \mathcal{L}_s^j) \quad (10)$$

Discussion about occlusion and cloth-changing. Since CSL and PSD mine the local similarity shared by different IDs, it may aggravate the negative impact of occlusion or cloth-changing. Below we briefly explain why our method is still valid. (1) We do not fuse local features (cascaded or weighted with global features) during inference, so local features with only occlusion information would not degrade performance. See Section 1 of the Appendix for details. (2) Since there is still a large gap between DG ReID and supervised ReID, the impact of cloth-changing can be ignored compared to improving generalization.

4. Experiments

4.1. Datasets and Evaluation Metrics

As shown in Table 2, we conduct experiments on four large-scale person re-identification datasets: Market1501[36], RandPerson[28], MSMT17[29] and CUHK03-NP[39]. For simplicity, we denote the datasets above as M, RP, MS, and C, respectively. We adopt the detected subset of the new protocol of CUHK03[12] (767 IDs for training and 700 IDs for evaluation), which is more challenging than the original CUHK03 protocol.

Table 1. Performance comparisons between ours and the state-of-the-art in single-source DG ReID on Market1501, RandPerson, MSMT17, and CUHK03-NP. Our results are highlighted in bold. The subscripts $_{50}$ and $_{152}$ denote using IBNNet$_{50}$ and IBNNet$_{152}$ as backbone, respectively.

Method	Reference	Training	Market		MSMT		CUHK03-NP	
			R1	mAP	R1	mAP	R1	mAP
SNR [10]	CVPR2020	Source: Market	55.1	33.6	-	-	-	-
CBN [41]	ECCV2020		91.3	77.3	25.3	9.2	-	-
OSNet-AIN [40]	TPAMI2021		94.2	84.4	23.5	8.2	-	-
QAConv [13]	ECCV2020		-	-	22.6	7.0	9.9	8.6
TransMatcher [14]	NeurIPS2021		-	-	47.3	18.4	22.2	21.4
QAConv+GS [15]	CVPR2022		91.6	75.5	45.9	17.2	19.1	18.1
MDA [18]	CVPR2022		-	-	33.5	11.8	-	-
PAT	Ours		**92.4**	**81.5**	**42.8**	**18.2**	**25.4**	**26.0**
SNR[10]	CVPR2020	Source: MSMT	70.1	41.4	-	-	-	-
CBN[41]	ECCV2020		73.7	45.0	72.8	42.9	-	-
QAConv [13]	ECCV2020		72.6	43.1	-	-	25.3	22.6
TransMatcher [14]	NeurIPS2021		80.1	52.0	-	-	23.7	22.5
QAConv+GS [15]	CVPR2022		79.1	49.5	79.2	50.9	20.9	20.6
MDA	CVPR2022		79.7	53.0	-	-	-	-
PAT	Ours		**72.2**	**47.3**	**75.9**	**52.0**	**24.2**	**25.1**
QACONV$_{50}$* [13]		Source: Multi	68.6	39.5	29.9	10.0	22.9	19.2
M^{3}L [34]			74.5	48.1	33.0	12.9	30.7	29.9
M^{3}L$_{IBN}$ [34]			75.9	50.2	36.9	14.7	33.1	32.1
RP Baseline [28]	ACMMM 20	Source: RandPerson	55.6	28.8	20.1	6.3	13.4	10.8
CBN [41]	ECCV2020		64.7	39.3	20.0	6.8	-	-
QAConv+GS [15]	CVPR2022		76.7	46.7	45.1	15.5	18.4	16.1
PAT	ours		**73.7**	**46.9**	**45.5**	**19.4**	**20.2**	**20.1**

Table 2. Statistics of Person ReID Datasets.

Dataset	# IDs	# images	# cameras
Market1501[36]	1,501	32,217	6
RandPerson[28]	8,000	1,801,816	-
MSMT17[29]	4,101	126,441	15
CUHK03-NP[39]	1,467	28,192	2

To evaluate the generalization of our models, we adopt a single-source protocol [13] and list the results of a multi-source protocol [34]. Under the setting of single-source, we use one dataset mentioned above for training (only the training set) and another one for testing (only the testing set). Under the multi-source protocol, one domain from multiple datasets is used for testing(only the testing set in this domain) and all the remaining domains are for training (only the training set). For evaluation metrics, the performance is evaluated quantitatively by mean average precision (mAP) and cumulative matching characteristic (CMC) at Rank-1 (R1), Rank-5 (R5), Rank-10 (R10).

4.2. Implementation Details

We use ViT-base with $stride = 16$ [4] pre-trained on ImageNet as our backbone (denoted as ViT-B/16 for short). The batch size is set to 64 and images are resized to 256×128. We adopt random flipping and local grayscale transformation [7] for data augmentation. To optimize the model, we use SGD optimizer with a weight decay of 10^{-4}. The learning rate increases linearly from 0 to 10^{-3} in the first 10 epochs then it decays in the following 50 epochs. The total training stage takes 60 epochs. For hyper-parameters, we conduct comprehensive experiments on the temperature parameter τ in section 4.4. Unless otherwise specified, we set α (the weight of similar categories), λ (the coefficient to balance soft label and hard label), and k (the number of clusters) to 0.5, 0.5, and 10, respectively. Besides, the label-smoothing parameter is 0.1. As for baseline, we use TransReID-B/16 [8] without SIE and JPM for a fair comparison. The training process contains 3 stages: (1) Extracting global and local features. (2) Performing CSL using local features (computing L_p). (3) Performing PSD using

Table 3. Improvements of our method (PAT) on different Transformers. The training set is Market.

Method	CUHK		MSMT	
	R1	mAP	R1	mAP
DeiT-Tiny[25]	9.1	9.5	14.2	4.5
PAT	**13.5**	**14.2**	**19.2**	**7.0**
DeiT-Small[25]	14.1	14.5	23.5	8.4
PAT	**18.4**	**18.2**	**27.1**	**10.0**
ViT-Small[4]	12.9	14.0	24.8	9.2
PAT	**15.1**	**15.2**	**26.9**	**10.0**
DeiT-base[25]	18.1	18.8	32.9	12.8
PAT	**20.3**	**21.0**	**36.2**	**14.9**
ViT-Base[4]	23.1	23.6	38.6	16.2
PAT	**25.4**	**26.0**	**42.8**	**18.2**

global features and soft labels generated by CSL (computing L_{tri} and L_s). The entire training process is end-to-end.

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

Single-source DG ReID To validate the performance of our model, we evaluate our framework on the single-source generalization ReID benchmark. Specifically, we use Market-train, MSMT-train, and RandPerson as the training sets, and use Market-test, MSMT-test, and CUHK03-test as the testing sets. Only the training set of the source datasets is used.

The experimental results are shown in Table 1. Our model outperforms the SOTA model under most settings and achieves a comparable performance with the SOTA model under the rest settings. In particular, under M→MS and M→C settings, our model's mAPs outperform the state-of-the-art model by 7.2% on average. When trained on Market, our model surpasses the SOTA model by 3.2% and 4.6% (test on CUHK03-NP) in R1 and mAP. When trained on MSMT, our model outperforms the SOTA by 1.6% and 7.2% (test on Market), 4.1% and 5.6% (test on MSMT) for R1 and mAP, respectively. This demonstrates the superiority of our model.

When trained on MSMT, our model still achieves SOTA on CUHK03-NP. Although TransMatcher [14] surpasses our method on Market, we experimentally found that the IBN [21, 9] trick greatly improved it, and our model could still surpass it if TransMatcher used a normal CNN as the backbone. This result is shown in the Appendix.

Large-scale DG ReID To further validate the generalization of our model, we also list the results under the multi-source protocol. The "Source: Multi" is the protocol proposed in M^{3}L [34] which we have introduced in section 4.1. Three of Market1501 , DukeMTMC-reID, MSMT17 and CUHK03-NP are used as source domains, and the rest as target domain under the multi-source protocol. Note that, since DukeMTMC-reID has been retracted, we utilize the large-scale publicly available dataset RandPerson as a substitute for multiple source domains in our training set.

As shown in Table 1, our method outperforms the SOTA model on RandPerson. Specifically, our model's mAP is 4.0% higher than the current best model (QAConv+GS) under the RP→C setting and reaches 56.9% under the RP→M setting. It's worth mentioning that when testing on MSMT, our method only requires a small dataset (such as Market) to outperform the results achieved by multiple source domains. This indicates that our method exhibits greater potential on small datasets.

Table 4. Ablation studies on the effectiveness of CSL and PSD. All methods are trained on Market and evaluated on MSMT.

Method	Market→MSMT			
	Rank1	Rank5	Rank10	mAP
Baseline (B)	38.6	52.1	58.1	16.2
B+SD [23]	35.6	49.7	56.4	14.6
B+CSL	40.2	53.7	59.6	17.0
B+CSL+PSD (ours)	**42.8**	**56.4**	**62.2**	**18.2**

4.4. Ablation Study

Improvement on Transformer. To investigate the improvement of our method on original Transformers (baseline), we conduct extensive experiments on different Transformers. As shown in Tab 3, our approach can improve the generalization of various Transformers on ReID task, especially when the model size is small. For example, our method surpasses baseline by 4.4% and 4.7% in R1 and mAP (Market → CUHK) when using DeiT-Tiny as the backbone.

Ablation study of main components of our model. To ensure that all components promote our model, we conduct an ablation study. All models are trained on Market and then tested on MSMT and CUHK03-NP, respectively. We choose the following models: (1) Baseline, namely TransReID-B/16 without SIE and JPM which is introduced in 4.1; (2) Baseline with conventional self-distillation (B + SD). We follow the self-distillation way designed for domain generalization [23]; (3) our model without PSD (B + CSL); (4) our model with all components, including CSL and PSD.

As shown in Table 4, Baseline + SD (self-distillation) results in a descent, which means that traditional self-distillation fails to improve the generalization of ReID. Since ReID is a fine-grained retrieval task, it is not suitable to use conventional self-distillation which usually requires

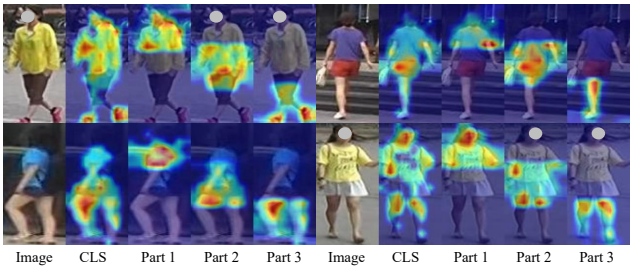

Figure 4. Attention map visualization. Original images are selected from the target domain (Market). We exhibit the visualizations of the class token and three part tokens.

Table 5. Ablation study for the number of parts.

Number of parts	M→C		M→MS	
	mAP	R1	mAP	R1
2	24.5	23.6	16.7	39.3
3	**26.0**	**25.4**	**18.2**	**42.8**
4	25.1	25.1	18.1	41.7
5	25.0	25.6	17.0	40.5

large inter-class distances. The results of (3) and (4) demonstrate our contribution. Firstly, our CSL module brings significant improvement. Secondly, PSD using the visual similarity of local parts to construct soft labels is effective.

Visualization of attention maps. To better understand the part tokens, we visualize the attention maps. The size of attention maps in each head is N×N (N is the number of patches). The attention maps of the class token and part tokens are resized to H×W (H and W denote the height and width of the input), then we turn them into heat maps with the original image. We fuse the attention results of the shallow layers, which contain more visual information than the deep ones. As shown in Figure 4, the class token mainly focuses on the whole images, while the "part tokens" pay attention to local areas like the upper body, legs, and backpacks. It shows that "part tokens" broaden the scope of attention, and provide more comprehensive multi-view information to the class token. Therefore, our model learns a more generic representation by utilizing local similarities.

Ablation study on the number of parts. In order to effectively capture visual similarity between different IDs in CSL, it's important to find the right balance in the number of partitions. If there are too few partitions (e.g., just one or two), it becomes difficult for different parts to exhibit visual similarity. Conversely, if there are too many partitions, there's a risk of clustering segments that lack semantic visual connections, leading to more noise. As shown in Table 5, dividing into three parts yields the best results, which is also intuitive.

Figure 5. Visualization of local features' ranking list in CSL.

Visualization of local features' ranking list. To verify whether CSL has mined local similarities, we show those samples that are closest to the current local feature, that is, the samples belonging to $\{\mathcal{K}_{p_i}^j\}_{j=1}^k$ in Section 3.2. The training set is Market. The red boxes represent randomly selected image tokens. As shown in Figure 5, the model can find samples with apparent similarity to the current sample in a specific area. For example, in the area attended by part token 1, not only people wearing dark red shirts but also backpacks were found. The above similarity is not dependent on the labels but completely derived from the data itself, which guides the model to learn generic features.

5. Conclusion

In this paper, we propose a pure Transformer-based framework (termed Part-aware transformer) for DG ReID for the first time. Specifically, we design a proxy task, named Cross-ID Similarity Learning (CSL), to mine local visual information shared by different IDs. This proxy task allows the model to learn generic features because it only cares about the visual similarity of the parts regardless of the ID labels, thus alleviating the side effect of domain-specific biases. Furthermore, we propose a part-guided self-distillation module to further improve the generalization of the global representation. Experimental results show that our method achieves state-of-the-art in DG ReID.

Acknowledgements

This study is supported by grants from National Key RD Program of China (2022YFC2009903/2022YFC2009900), the National Natural Science Foundation of China (Grant No.62122018, No.62020106008, No.61772116, No.61872064), Fok Ying-Tong Education Foundation (171106), and SongShan Laboratory YYJC012022019.

References

[1] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[2] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[3] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[6] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[7] Yunpeng Gong. A general multi-modal data learning method for person re-identification. *arXiv preprint arXiv:2101.08533*, 2021.

[8] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[9] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. *arXiv preprint arXiv:1905.03422*, 2019.

[10] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.

[12] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[13] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *Proceedings of the IEEE/CVF European conference on computer vision*. Springer, 2020.

[14] Shengcai Liao and Ling Shao. Transmatcher: Deep image matching through transformers for generalizable person re-identification. *Advances in Neural Information Processing Systems*, 34, 2021.

[15] Shengcai Liao and Ling Shao. Graph sampling based deep metric learning for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[16] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021.

[17] Xinyu Lyu, Lianli Gao, Pengpeng Zeng, Heng Tao Shen, and Jingkuan Song. Adaptive fine-grained predicates learning for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[18] Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, and Heng Tao Shen. Meta distribution alignment for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[19] Hao Ni, Jingkuan Song, Xiaosu Zhu, Feng Zheng, and Lianli Gao. Camera-agnostic person re-identification via adversarial disentangling learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[20] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF European conference on computer vision*. Springer, 2016.

[21] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Asian Conference on Computer Vision*, 2018.

[22] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[23] Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan, and Fahad Shahbaz Khan. Self-distilled vision transformer for domain generalization. *arXiv preprint arXiv:2207.12392*, 2022.

[24] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 2019.

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 2021.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[27] Yufei Wang, Haoliang Li, Lap-pui Chau, and Alex C Kot. Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[28] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for

generalizable person re-identification. In *Proceedings of the 28th ACM international conference on multimedia*, 2020.

[29] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[30] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[31] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Pattern Recognition*.

[32] Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[33] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[34] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[35] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015.

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[37] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[38] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017.

[39] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[40] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44, 2021.

[41] 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 *Proceedings of the IEEE/CVF European conference on computer vision*. Springer, 2020.