Visual Intelligence

RESEARCH Open Access

Learning convolutional multi-level transformers for image-based person re-identification

Peilei Yan[1], Xuehu Liu[2], Pingping Zhang[1*] and Huchuan Lu[2]

Abstract

As a vital vision task, person re-identification (Re-ID) aims to retrieve the same person under non-overlapping cameras. It is a very challenging task due to the presence of complex backgrounds, diverse illuminations and different perspectives. In this work, we integrate the advantages of convolutional neural networks (CNNs) and transformers, and propose a novel learning framework named convolutional multi-level transformer (CMT) for image-based person Re-ID. More specifically, we first propose a scale-aware feature enhancement (SFE) module to extract multi-scale local features from a pre-trained CNN backbone. Then, we introduce a part-aware transformer encoder (PTE) to further mine discriminative local information guided by global semantics. Finally, a deeply-supervised learning (DSL) technique is adopted to optimize the proposed CMT and improve its training efficiency. Extensive experiments on four large-scale Re-ID benchmarks demonstrate that our method performs favorably against several state-of-the-art methods.

Keywords: Person re-identification (Re-ID), Vision transformer, Global-local features, Deeply-supervised learning (DSL)

1 Introduction

Person re-identification (Re-ID) aims to retrieve specific persons in a scene based on the content of images or videos taken at different times and places. It has drawn much attention due to its diversified real-world applications, such as safe communities, intelligent surveillance and criminal investigations [1–3]. Although great success has been achieved, there are still many challenges in person Re-ID, such as object occlusion, illumination change, pose distortion and background clutter.

In the past two decades, great progresses have been achieved in the typical image-based Re-ID task [4]. The accomplishment of this task largely depends on the robust representations of person images. In fact, early person Re-ID methods [5–7] primarily focus on the hand-crafted feature extraction and the similarity metric design. With the development of deep learning technologies, many works focus on the end-to-end learning of more discriminative features by designing complex deep convolutional neural networks (CNNs). In addition, local information is also discriminative and helpful in retrieving the target person. As illustrated in the upper row of Fig. 1, the features extracted by the CNN backbone are horizontally divided into multiple parts in such part-level feature extraction methods as part-based convolutional baseline (PCB) and research has demonstrated that the PCB method has achieved significant performance improvements [8]. However, the convolutional layers usually model the relationship between pixels in a small neighborhood and cannot realize the global modeling of person images. Thus, most CNN-based methods [9–11] are ineffective when facing certain challenges such as varied posture, occlusion, and background clutter.

*Correspondence: zhpp@dlut.edu.cn
[1] School of Artificial Intelligence, Dalian University of Technology, Dalian, 116024, China
Full list of author information is available at the end of the article

Recently, transformers [12] have achieved excellent performance in natural language processing and computer vision. The key reason is that transformers are global operations based on self-attention and can model the relationship between all input elements. As a result, several attempts have been made to accomplish person Re-ID using transformers. For example, Zhu et al. [13] introduced an auto-aligned structure and enhanced the ability of transformers to extract more discriminative features. He et al. [14] proposed a pure transformer architecture to integrate camera and viewpoint information and achieved excellent performance in object re-identification. Although effective, these transformer-based methods require a large number of transformer blocks, resulting in high model complexity. In addition, these works seldom take into account the local information of persons, which is crucial for person Re-ID. Therefore, there is still much room for improvement in current transformer-based methods.

In this work, we take advantage of CNNs and transformers, and propose a novel learning framework named convolutional multi-level transformer (CMT) for image-based person Re-ID. More specifically, we first utilize a scale-aware feature enhancement (SFE) module to extract multi-scale local features from deep CNN backbones. As a result, they can capture multi-granularity representations of various appearances in person images. Then, we introduce a part-aware transformer encoder (PTE) to further extract local discriminative information guided by global semantics. As shown in the bottom row of Fig. 1, we incorporate the idea of feature partitioning into the transformer and design a recursive transformer structure. This structure can generate hierarchical features for diverse local parts, resulting in great performance improvements. Finally, we adopt a deeply-supervised learning (DSL) technique to optimize the proposed CMT and improve its training efficiency. Extensive experiments on four large-scale Re-ID benchmarks demonstrate that our method performs favorably against most state-of-the-art methods.

The main contributions are summarized as follows:

1) A novel global-local feature learning framework (i.e., CMT) is proposed for robust person Re-ID.
2) A SFE module is proposed to extract multi-scale local features, capturing multi-granularity representations of person images.
3) A PTE is proposed to further extract local discriminative information guided by global semantics. The PTE can generate hierarchical features for diverse local parts.
4) Extensive experiments demonstrate that our proposed framework can effectively extract robust and discriminative features. It achieves state-of-the-art performances on four large-scale Re-ID benchmarks.

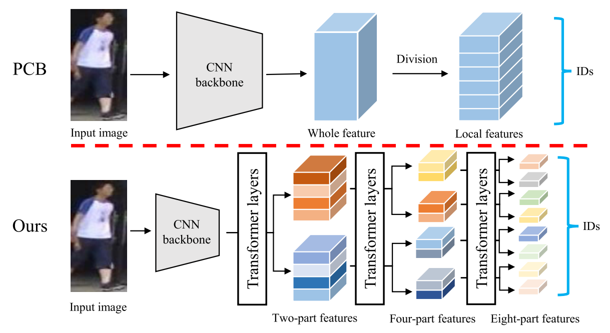

Figure 1 The insight of our proposed CMT. Upper row: Previous horizontal part divisions in models such as PCB [8]; Bottom row: Our hierarchical and recursive part encoding with transformers

2 Related works

2.1 Part-based person re-identification

In recent years, image-based person Re-ID has achieved great improvements in performance. Generally, existing person Re-ID methods mainly focus on extracting discriminative global features from entire images. However, focusing merely on the global information of persons has some limitations, such as ignoring the effectiveness of local cues. Fine-grained local part features such as T-shirt and black backpack can be very useful to identify persons in complex scenes. As a typical practice, many researchers resort to part features for pedestrian image description. In particular, Sun et al. [8] proposed a method to divide spatial features into horizontal strips to improve the Re-ID performance. Wang et al. [10] utilized a multi-branch network to extract the multi-granularity features of persons. Zheng et al. [15] proposed a coarse-to-fine pyramid model to fuse global and local features. Yang et al. [16] designed a patch-wise loss function to guide the effective learning of patch features. Cho et al. [17] leveraged the complementary relationships between global and local features to refine the pseudo labels of parts and reduce label noises. Different from the above local-based methods, we propose a recursive structure to iteratively mine local features under global semantic guidance. By hierarchical learning, our method can generate diverse local part features of individual persons, resulting in sufficient richness of image information and robustness of the Re-ID results.

2.2 Attention-based person re-identification

Visual attention mechanisms aim to highlight relevant information and suppress irrelevant information. Inspired by the advantages of attention mechanisms, researchers have proposed various attention-based methods to extract distinguishable features for person Re-ID. For example, Chen et al. [18] proposed a mixed high-order attention to capture the subtle differences among pedestrians. Rao et al. [19] presented a counterfactual attention to capture

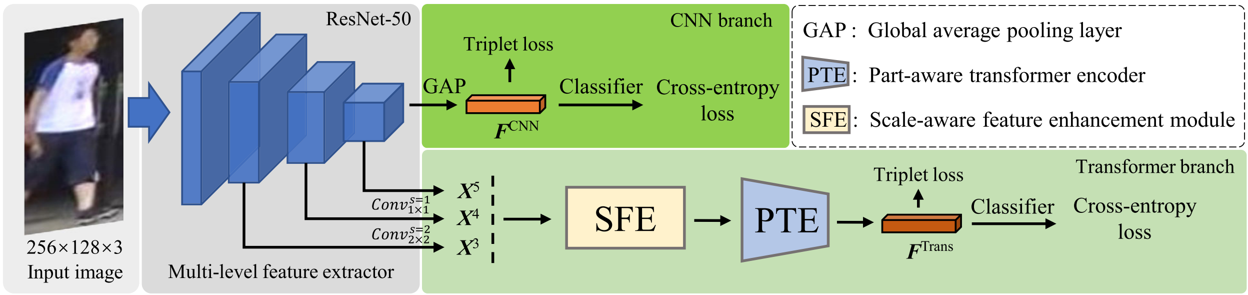

Figure 2 The architecture of our proposed framework. Given a person image, we first utilize the multi-level feature extractor (MFE) to extract multi-level feature maps. Then, for the CNN branch, we employ a GAP layer to obtain the CNN feature. For three stages of the MFE, we introduce the transformer branch, and utilize the SFE module and PTE to further extract multi-scale and diverse part features, respectively. All these branches are supervised by the triplet loss and cross-entropy loss for model training

more discriminative representations. Chen et al. [20] built a pyramid attention to explore attentive regions in a multi-scale manner. Zhang et al. [21] proposed a relation-aware attention to capture the global structural information from persons. Li et al. [22] presented a harmonious attention to reduce the misalignments of the same persons. Different from the above attention-based methods, we introduce attention mechanisms to capture long-range dependencies between local features, leading to much better results.

2.3 Transformer-based person re-identification

In fact, transformers [12] are initially proposed for processing sequential data. With the global modeling ability, transformers have been recently introduced to many computer vision tasks, including person Re-ID. For image-based person Re-ID, He et al. [14] first utilized a pure transformer-based structure [23] to learn discriminative features. Zhu et al. [13] added learnable vectors of part tokens to learn part features and integrated part alignments into the self-attention. Lai et al. [24] utilized transformers to achieve adaptive part divisions. Li et al. [25] introduced a diverse part discovery with part-aware transformers for occluded person Re-ID. Liao and Shao [26] built a transformer-based deep image matching for generalizable person Re-ID. Wang et al. [27] proposed a self-guided transformer framework to explore the relations of body parts for feature alignment. Chen et al. [28] proposed an omni-relational high-order transformer for person Re-ID. Ma et al. [29] proposed a pose-guided transformer to mine the inter-part and intra-part relations for occluded person Re-ID. Liu et al. [30] designed a trigeminal transformer to simultaneously encode the spatial, temporal and spatial-temporal features in complex videos. These transformer-based methods have achieved superior performances. However, they generally lack desirable local properties. Different from them, we introduce a hybrid structure combining CNNs and transformers for more effective person Re-ID.

3 Proposed method

As illustrated in Fig. 2, the proposed framework mainly includes three key modules: a multi-level feature extractor (MFE), the SFE module and a PTE. More specifically, the MFE utilizes a pre-trained CNN backbone (e.g., ResNet-50 [31]) to extract multi-level features of person images. Afterwards, the SFE module adopts multi-scale dilated convolutions [32] with residual connections to capture multi-granularity feature representations. Furthermore, with a hierarchical structure, PTE further mines local discriminative information guided by global semantics. Finally, the DSL technique is utilized to optimize the whole framework. We will elaborate on these key components in the following subsections.

3.1 Multi-level feature extractor

As illustrated in the left part of Fig. 2, we utilize the ResNet-50 [31] pre-trained on ImageNet to extract multi-level features. Similar to previous works [8, 10, 33], we remove the fully-connected layers after the global average pooling (GAP) layer, and change the stride of the fifth stage to 1, resulting in a 1/16 feature resolution of input images. In addition, we take the outputs of stages 3, 4 and 5, and introduce an additional convolutional layer to generate size-fixed multi-level features.

3.2 Scale-aware feature enhancement

Due to the variations of persons in scenes, multi-scale information [33] is effective for robust appearance representations. Thus, we propose the SFE module to extract multi-scale features at three stages of the backbone network.

The structure of our proposed SFE module is illustrated in Fig. 3. Given an input $\boldsymbol{X}^i$ ($i = 3, 4, 5$), we first reduce the channel numbers to a quarter of $\boldsymbol{X}^i$ by a convolutional layer and obtain $\tilde{\boldsymbol{X}}^i$. Then, we utilize four dilated convolutional layers to generate multi-scale features and gradually ex-

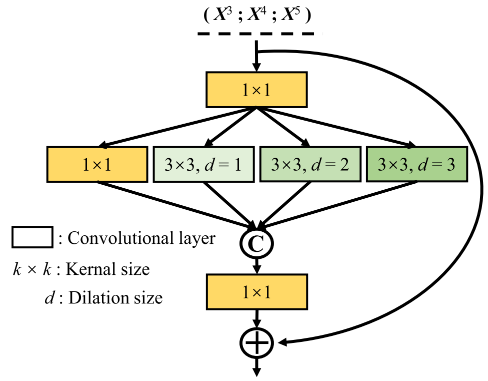

Figure 3 The structure of our scale-aware feature enhancement

tend the receptive fields [32].

$$\begin{aligned} \boldsymbol{M}_1 &= \mathrm{Conv}_1\left(\tilde{\boldsymbol{X}}^i\right), & \boldsymbol{M}_2 &= \mathrm{Conv}_2\left(\tilde{\boldsymbol{X}}^i\right), \\ \boldsymbol{M}_3 &= \mathrm{Conv}_3\left(\tilde{\boldsymbol{X}}^i\right), & \boldsymbol{M}_4 &= \mathrm{Conv}_4\left(\tilde{\boldsymbol{X}}^i\right). \end{aligned} \tag{1}$$

Then, they are concatenated in the channel and aggregated by another convolutional layer. Meanwhile, a residual connection is utilized to obtain the final output of SFE,

$$\boldsymbol{Y}^i = \boldsymbol{X}^i + \mathrm{Conv}\left([\boldsymbol{M}_1; \boldsymbol{M}_2; \boldsymbol{M}_3; \boldsymbol{M}_4]\right), \tag{2}$$

where [;] means the concatenation in the channel. In fact, due to the utilization of different kernel sizes and dilation sizes, our SFE module is able to capture multi-scale local cues for scale-aware feature enhancement.

3.3 Part-aware transformer-based encoder

In addition to SFE, we employ PTE to further extract partware fine-grained representations with transformers. As illustrated in Fig. 4, our PTE is designed with a recursive and hierarchical structure, which progressively generates diverse part features with global semantic guidance.

Formally, the PTE takes $\boldsymbol{Y}^i$ as input and introduces hierarchical divisions for diverse part features. It should be noted that all the transformers at the same stage share weights for computation reduction. The structure of the transformers is identical to [23]. At the 2^k-part learning stage ($k = 1, 2, \ldots$), we first use a 1×1 convolutional layer to halve the number of channels. Then, we reshape the feature map into a sequence representation $\boldsymbol{F}_{2^k} \in \mathbb{R}^{HW \times C}$. Here, H and W denote the height and width of the input image, respectively. C represents the number of channels. The class token $\boldsymbol{F}^{\mathrm{cls}}_{2^{k-1}} \in \mathbb{R}^{1 \times C}$ from the 2^{k-1}-part learning stage is concatenated into the sequence to guide the fine-grained features. In addition, a new class token $\boldsymbol{F}^{\mathrm{cls}}_{2^k} \in \mathbb{R}^{1 \times C}$ is also concatenated into the sequence to summarize contextual information. Finally, the position embedding $\boldsymbol{F}^{\mathrm{pos}}_{2^k} \in \mathbb{R}^{(HW+2) \times C}$ is added to the sequence. For the 2^k-part learning stage, the input embedding for the j-th part transformer is:

$$\tilde{\boldsymbol{F}}_{2^k,j} = \left[\boldsymbol{F}^{\mathrm{cls}}_{2^k,j}; \phi\left(\boldsymbol{F}^{\mathrm{cls}}_{2^{k-1},n}\right); \tilde{\boldsymbol{F}}_{2^{k-1},j}\right] + \boldsymbol{F}^{\mathrm{pos}}_{2^k,j}, \tag{3}$$

where $j \in \{1, 2, \ldots, 2^k\}$, and n is equal to $j/2$ when j is even; otherwise n is equal to $(j+1)/2$. ϕ is a linear projection to align the channel numbers of features. The above input goes through several transformer layers, each of which includes a multi-head self attention (MHSA) module and a feed forward network (FFN). After building the hierarchical structure, we generate the part features as:

$$\boldsymbol{F}^p = \left[\tilde{\boldsymbol{F}}^{\mathrm{cls}}_{2^k,1}; \tilde{\boldsymbol{F}}^{\mathrm{cls}}_{2^k,2}; \ldots; \tilde{\boldsymbol{F}}^{\mathrm{cls}}_{2^k,2^k}\right]. \tag{4}$$

From the above equations and Fig. 4, one can see that our proposed PTE uses transformers to generate hierarchical local features with the guidance of global semantics. This recursive and hierarchical design can not only generate multi-scale and multi-granularity features but also provide global guidance for more discriminative features, enhancing the extraction of local features. In addition, we apply transformers to extract local features and stack fewer transformer blocks, which can significantly reduce the model complexity.

3.4 Deeply-supervised learning

As illustrated in Fig. 2, we utilize both the feature F^{CNN} generated from the CNN branch and the features F^{Trans} from the transformer branch for inference. To train the whole framework, we adopt the DSL technique [33, 34], which makes the network optimization a task that is easy to complete. At each branch, we use the label-smoothed cross-entropy loss [35] and the batch-hard triplet loss [36]. The label-smoothed cross-entropy loss is defined as:

$$\mathcal{L}_{\mathrm{ce}} = \sum_{i=1}^{N} -q_i \ln(p_i), \tag{5}$$

where p_i is the predicted logit of identity i and q_i is the ground-truth label. The batch-hard triplet loss is defined as:

$$\mathcal{L}_{\mathrm{tri}} = [d_{\mathrm{pos}} - d_{\mathrm{neg}} + m]_+, \tag{6}$$

where d_{pos} and d_{neg} are defined as the distance of positive sample pairs and negative sample pairs, respectively. $[x]_+$ is $\max(0, x)$ and m is the distance margin.

Finally, the overall loss can be summarized as:

$$\mathcal{L}_{\mathrm{all}} = \mathcal{L}_{\mathrm{ce}} + \mathcal{L}_{\mathrm{tri}} + \lambda \sum_{k=1}^{K} \left(\mathcal{L}^k_{\mathrm{ce}} + \mathcal{L}^k_{\mathrm{tri}}\right), \tag{7}$$

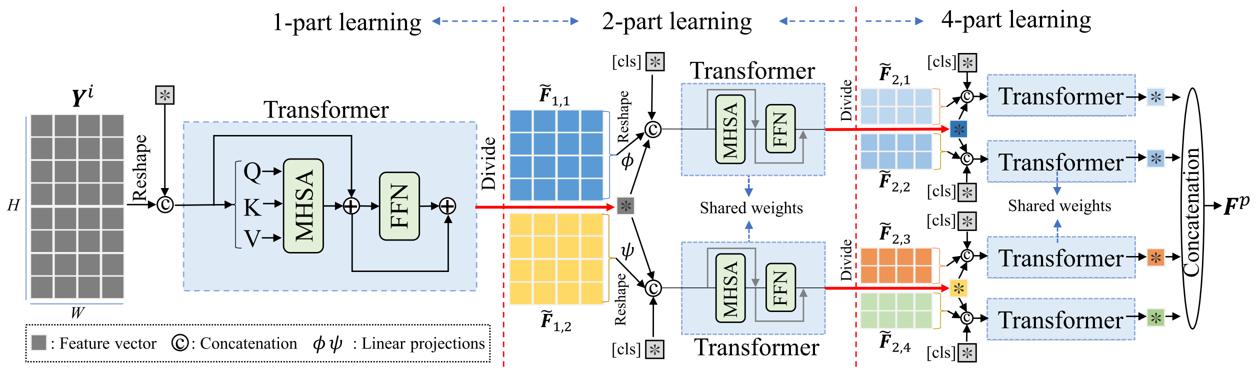

Figure 4 Illustration of our proposed part-aware transformer-based encoder. *H* and *W* denote the height and width of the input image, respectively. Q, K, and V represent query, key and value, respectively. MHSA represents a multi-head self attention and FFN means a feed forward network

where K is the number of stages. λ is the balanced coefficient for the multiple loss terms.

4 Experiments

4.1 Datasets and evaluation metrics

We conducted extensive experiments on four widely-used person Re-ID datasets, i.e., Market1501 [37], DukeMTMC-ReID [38], CUHK03-NP [39] and MSMT17 [40]. The Market1501 was collected from six cameras and has 1501 pedestrians (751 for training and 750 for testing). The DukeMTMC-ReID was collected from eight cameras with 1404 pedestrians (702 for training and 702 for testing). The CUHK03-NP dataset consists of 1467 pedestrians, which are divided into two sub-datasets: one with manual labeling and the other with bounding boxes labeled by a person detector. The MSMT17 is a large-scale dataset deriving from 15 cameras with 4101 pedestrians (1041 for training and 3010 for testing). Table 1 provides more detailed statistics of the four datasets. Following previous works [4, 33], we compute the mean average precision (mAP) and cumulative matching characteristics (CMC) at rank-1 for performance evaluation.

Table 1 Statistics of our used datasets

Dataset	ID	Image	Train	Test	#Cameras
Market1501	1501	32668	12936	19732	6
DukeMTMC-ReID	1404	36411	16522	19889	8
CUHK03-NP-Labeled	1467	14096	7368	6728	10
CUHK03-NP-Detected	1467	14096	7365	6732	10
MSMT17	4101	126441	32621	93820	15

4.2 Implementation details

In this work, all the experiments are performed with the PyTorch toolbox[1] and one GeForce RTX 3090 GPU. We utilize the ResNet-50 pre-trained on ImageNet as our backbone. In addition, we balance the accuracy and complexity, and ultimately choose to extract four parts through the PTE. To extract the multi-scale features by the SFE module, a 1 × 1 convolutional layer and three 3 × 3 dilated convolutional layers are used to gradually extend the receptive fields. Dilation sizes d are 1, 2 and 3, respectively. During training, all images of pedestrians are resized to 256 × 128 and augmented by random cropping, horizontal flipping and random erasing [41]. In one mini-batch, 16 identities are randomly sampled and each identity has 4 images. The Adam optimizer [42] is deployed with an initial learning rate of 3.5×10^{-4}, which is multiplied by 0.4 every 20 epochs until 180 epochs. The source code is released at https://github.com/AI-Zhpp/CMT.

4.3 Comparison with state-of-the-art methods

In this subsection, we compare our method with other state-of-the-art methods. The comparison results on four public Re-ID benchmarks are presented in Table 2. The detail analysis is as follows:

Market1501 As for CNN-based methods, PCB [8] and MGN [10] mine diverse part features by horizontal strip features and reach 81.6% mAP and 86.9% mAP on Market1501, respectively, which validate the reasonableness of part learning in Re-ID. In our method, we adopt a hierarchical transformer-based structure to progressively extract multi-granularity part representations. Thus, our method achieves the best mAP and outperforms PCB and MGN by 8.3% and 3.0%, respectively. Even in comparison with transformer-based methods, such as AAformer [13], TransReID [14], APD [24] and HAT [33], our method still delivers a better performance.

DukeMTMC-ReID On this dataset, our method shows superior performances. The mAP and rank-1 accuracy are

[1] http://pytorch.org.

Table 2 Performance(%) comparison with state-of-the-arts. The best performance is marked in bold and the second-best performance is underlined. * indicates that the methods are using camera information

Methods	Backbones	Market1501		DukeMTMC-ReID		CUHK03-NP				MSMT17	
						Labeled		Detected			
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
DuATM [44]	DenseNet121	76.6	91.4	64.6	81.8	–	–	–	–	–	–
Mancs [45]	ResNet50	82.3	93.1	71.8	84.9	63.9	69.0	60.5	65.5	–	–
IANet [46]	ResNet50	83.1	94.4	73.4	83.1	–	–	–	–	46.8	75.5
BoT [47]	ResNet50	85.7	94.1	75.9	86.2	73.8	74.7	71.2	73.4	49.8	74.0
PCB [8]	ResNet50	81.6	93.8	69.2	83.3	–	–	57.5	63.7	40.4	68.2
SPReID [48]	ResNet152	83.4	93.7	73.3	85.9	–	–	–	–	–	–
AANet [49]	ResNet152	83.4	93.9	74.3	87.7	–	–	–	–	–	–
CASN [50]	ResNet50	82.8	94.4	73.7	87.7	68.0	73.7	64.4	71.5	–	–
CAMA [51]	ResNet50	84.5	94.7	72.9	85.8	–	–	64.2	66.6	–	–
BATNet [52]	ResNet50	84.7	95.1	77.3	87.7	76.1	78.6	73.2	76.2	56.8	79.5
MHN-6 [18]	ResNet50	85.0	95.1	77.2	89.1	72.2	77.2	65.4	71.7	–	–
BFE [53]	ResNet50	86.2	95.3	75.9	88.9	76.7	79.4	73.5	76.4	51.5	78.8
MGN [10]	ResNet50	86.9	95.7	78.4	88.7	67.4	68.0	66.0	68.0	–	–
ABDNet [11]	ResNet50	88.3	95.6	78.6	89.0	–	–	–	–	60.8	82.3
Pyramid [15]	ResNet101	88.2	95.7	79.0	89.0	76.9	78.9	74.8	78.9	–	–
JDGL [54]	ResNet50	86.0	94.8	74.8	86.6	–	–	–	–	52.3	77.2
OSNet [9]	OSNet	84.9	94.8	73.5	88.6	–	–	67.8	72.3	52.9	78.7
SNR [55]	ResNet50	84.7	94.4	73.0	85.9	–	–	–	–	–	–
SCSN [43]	ResNet50	88.5	95.7	79.0	**91.0**	–	–	–	–	–	–
ISP [56]	HRNet48	88.6	95.3	80.0	89.6	74.1	76.5	71.4	75.2	–	–
HAA [57]	ResNet50	89.5	95.8	80.4	89.0	–	–	–	–	–	–
CDNet [58]	CDNet	86.0	95.1	76.8	88.6	–	–	–	–	54.7	78.9
APNet [20]	ResNet50	89.0	**96.1**	78.8	89.3	**81.1**	**83.5**	78.1	80.9	59.0	80.8
AAformer [13]	ViT-B/16	87.7	95.4	80.0	90.1	77.8	79.9	74.8	77.6	62.6	83.1
TransReID* [14]	ViT-B/16	88.2	95.0	80.6	89.6	–	–	–	–	**64.9**	**83.3**
APD [24]	ResNet50	87.5	95.5	74.2	87.1	73.8	77.0	70.6	74.6	57.1	79.8
HAT [33]	ResNet50	89.5	95.6	81.4	90.4	80.0	82.6	75.5	79.1	61.2	82.3
CMT (Ours)	ResNet50	**89.9**	95.8	**82.1**	90.5	80.7	82.9	**78.4**	**81.6**	63.5	**83.3**

82.1% and 90.5%, respectively, and exceed most of the current methods. It is noted that SCSN [43] integrates salient features using a cascaded network architecture, resulting in a rank-1 accuracy of 91%. Different from it, our method takes advantages of CNNs and transformers to incorporate global and local features. Compared with SCSN, our method gains a 3.1% improvement in mAP.

CUHK03-NP On two sub-datasets of CUHK03-NP, our method consistently achieves competitive results. Meanwhile, APNet [20] utilizes a pyramid attention to explore the discriminative regions of person images, and achieves 81.1% mAP and 78.1% mAP on the labeled and detected sub-datasets of CUHK03-NP, respectively. Different from APNet, our method extracts fine-grained partial features by multi-stage transformers. Compared with APNet, our method improves the mAP on the detected CUHK03-NP by 0.3%.

MSMT17 On this dataset, our framework also attains comparable performance in terms of mAP and rank-1. In fact, TransReID achieves the best mAP and rank-1 on MSMT17. However, TransReID uses ViT [23] as the backbone to capture long-range dependencies, which consumes high cost complexity and extremely impacts the inference speed. In contrast, our method uses ResNet-50 to extract local representations and combines part-aware transformers for fine-grained cues. Thus, our method attains a significant improvement of efficiency over TransReID. In addition, TransReID utilizes camera information for performance boosting, while our method does not utilize camera information but unifies the strengths of CNNs and transformers, which leads to the second-best performance on MSMT17.

Model complexities To further clarify the computation advantages, we compare the model complexity of some typical methods in Table 3. We use floating point operations per second (FLOPs) to test our model's computational complexity. As can be seen in Table 3, our proposed model shows great advantages over other transformer-based methods in terms of FLOPs. We also note that our proposed model has more parameters. This problem

Table 3 Comparisons of model complexities. Both CNN-based methods and transformer-based methods are selected for comparisons. Params means parameter. FLOPs denotes floating point operations per second

Methods	Backbones	Market1501		Params. (M)	FLOPs (G)
		mAP(%)	Rank-1(%)		
BoT [47]	ResNet50	85.7	94.1	25.64	4.08
ABDNet [11]	ResNet50	88.3	95.6	53.64	6.27
APNet [20]	ResNet50	89.0	96.1	29.90	8.16
HAT [33]	ResNet50	89.5	95.6	219.44	21.44
TransReID* [14]	ViT-B/16	88.2	95.0	104.71	178.52
CMT (Ours)	ResNet50	**89.9**	**95.8**	286.54	21.32

Table 4 Ablation analysis of key modules. Params means parameter. FLOPs denotes floating point operations per second

Methods	MSMT17		Params. (M)	FLOPs (G)
	mAP(%)	Rank-1(%)		
Baseline	49.8	74.0	25.64	4.08
+ PTE	62.6	82.4	238.79	16.01
+ SFE	63.5	83.3	286.54	21.32

can be solved by light-weight designs. CNN-based methods generally have fewer parameters and FLOPs. However, their performances are usually worse than those of transformer-based methods. Overall, our proposed model achieves a good balance between the Re-ID performance and the model complexity.

4.4 Ablation studies

To verify the effectiveness of our proposed modules, we conduct ablation experiments on the MSMT17 dataset.

Effectiveness of key modules The ablation results of our key modules are reported in Table 4. For the baseline method, we fine-tune ResNet-50 on MSMT17 and adopt GAP to obtain a feature vector for testing, which achieves 49.8% mAP and 74.0% rank-1 accuracy. Then, we add our PTE to the baseline to further extract diverse part features at three stages. In PTE, the global feature is recursively passed into part-aware transformers and used to refine part features. Thus, our PTE brings significant improvements over the baseline (i.e., 12.8% mAP and 8.4% rank-1 accuracy). Furthermore, we insert SFE to enhance the local features before PTE. SFE can capture multi-granularity representations of person images. Therefore, it brings performance improvement (i.e., 0.9% mAP and 0.9% rank-1 accuracy). Overall, the resulting improvements verify the effectiveness of our SFE module and PTE, which play a critical role in the extraction of multi-scale and discriminative features.

Table 5 Ablation analysis of the PTE module. Params means parameter. FLOPs denotes floating point operations per second

Methods	#Parts	MSMT17		Params. (M)	FLOPs (G)
		mAP(%)	Rank-1(%)		
Baseline	–	49.8	74.0	25.64	4.08
+ PTE	1	56.7	79.6	127.33	11.38
	2	62.0	82.1	192.19	14.50
	4	62.6	82.4	238.79	16.01

Table 6 Ablation results of deploying PTE after different levels of ResNet-50

Methods	mAP(%)	Rank-1(%)
ResNet-50	49.8	74.0
Res3 + PTE	56.7	77.2
Res4 + PTE	59.5	80.8
Res5 + PTE	53.0	76.9
Res3, Res4 + PTE	61.8	81.1
Res3, Res4, Res5 + PTE	62.6	82.4

Effects of the PTE module In PTE, we introduce a hierarchical transformer to split and encode part features. The ablation results are summarized in Table 5. As the recursive and hierarchical structure advances, the accuracy also achieve significant improvements. It can be observed that the mAP and rank-1 accuracy are improved by 5.3% and 2.5%, repectively when the spatial features are divided into two parts. With the increase of the part numbers, the diversity of fine-grained clues is captured. In our work, four parts are extracted for the trade-off between accuracy and complexity,.

Effects of PTE at different levels In experiments, we deploy PTE at different levels of ResNet-50 to realize multi-level part learning. The experimental results are listed in Table 6. From the results, one can observe that the deployment of PTE at a single level can improve performance. The best performance is achieved when PTE is deployed at three levels of ResNet-50. This fact confirms that multi-level representation learning is helpful to achieve better performances of person Re-ID.

Effectiveness of DSL In this work, we introduce the DSL for better model training. By deploying losses at different stages, the ablation results are reported in Table 7. It can be observed that the single deployment of supervision at the 4-part learning stage is not sufficient and more supervision is needed to train the entire framework well. When supervision is deployed at all stages, we can obtain the best performance.

Effects of different transformer layers and attention heads The number of transformer layers and attention heads may change the structure and performance of our PTE. Thus, we perform ablation experiments to examine the effects

Table 7 Ablation results of DSL

Methods	mAP(%)	Rank-1(%)
4-part	46.0	70.0
4-part + 1-part	60.4	81.5
4-part + 2-part	50.3	72.4
4-part + 2-part + 1-part	63.5	83.3

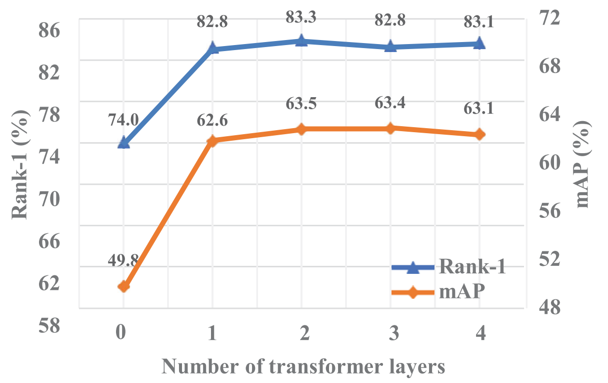

Figure 5 Ablation results with different transformer layers

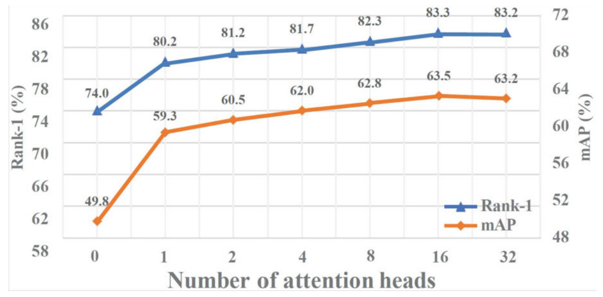

Figure 6 Ablation results with different attention heads

of transformer layers and attention heads. As shown in Fig. 5, the performance of our proposed model is significantly reduced without transformer layers. The performance degradation indicates that the features obtained solely from CNNs are not robust enough, and transformers can implicitly learn more discriminative information. In addition, we observe that when the number of transformer layers is set to 2, the best performance can be achieved. Meanwhile, with the increase of transformer layers, there are some fluctuating changes in performance. This may be because different transformer layers can change the local features. Furthermore, from Fig. 6, it can be observed that as the number of attention heads increases, the retrieval accuracy continues to be improved. Nevertheless, the performance is saturated when the number of attention heads is equal to 16. Based on the aforementioned facts, we set the numbers of transformer layers and attention heads to 2 and 16, respectively.

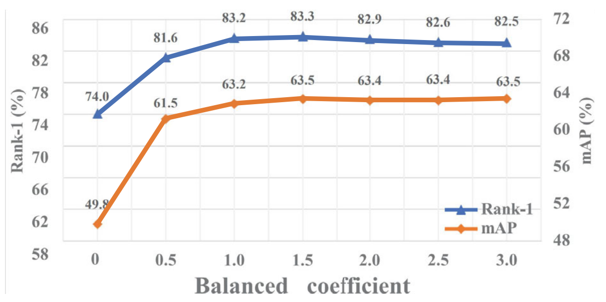

Figure 7 Ablation results with different balanced coefficient λ

Effects of the balance coefficient λ In our work, we utilize λ to balance different loss terms in Eq. (7). To verify its effect, we conduct experiments by changing the coefficient λ from 0 to 3. As displayed in Fig. 7, with the increase of λ, the performance continues to be improved. When λ is set to 1.5, the best performance can be achieved.

4.5 Visualization analysis

Visualization of feature maps To verify the effectiveness of the proposed modules, we further visualize the features of person examples. The visualizations are shown in Fig. 8. In each example, from left to right, there are the original image, baseline features, SFE features, and PTE features. It can be observed that increasingly detailed information is captured with the gradual utilization of our key modules. Moreover, the feature maps obtained from the baseline generally focus on salient regions, such as the heads or shoes of persons. With the utilization of the SFE module to extract multi-scale features, our model can capture more meaningful information, such as bags and clothing. With the utilization of the PTE module to extract diverse local features, our model can capture more detailed information, such as torso details. The visualization results demonstrate that our PTE can indeed mine discriminative and diverse local cues guided by global semantics. The visualizations intuitively verify the effectiveness of our proposed SFE module and PTE.

Meanwhile, we visualize the different parts in PTE for qualitative comparison in Fig. 9. Comparing the 2-nd, 3-rd and 4-th columns, it can be observed that more local cues can be captured as the number of parts increases. These visualization comparisons further explain the reasonableness of our PTE.

Retrieval results We also visualize the retrieval results on the MSMT17 dataset in Fig. 10. It can be observed that the retrieval accuracies are improved when the proposed key modules are gradually added to the baseline method.

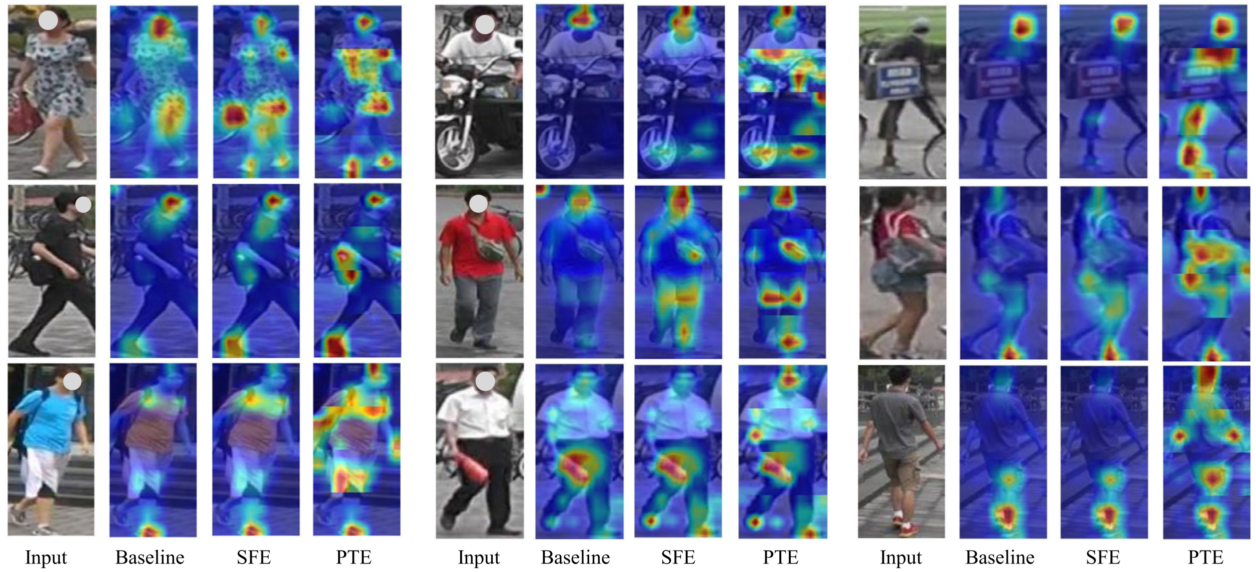

Figure 8 Visualizations of features obtained from Baseline, SFE and PTE

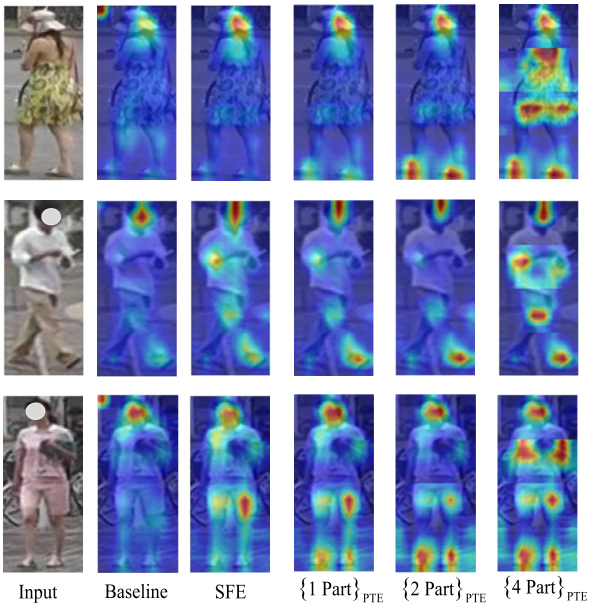
Input
Baseline
SFE
$\{1\text{ Part}\}_{\text{PTE}}$
$\{2\text{ Part}\}_{\text{PTE}}$
$\{4\text{ Part}\}_{\text{PTE}}$

Figure 9 Visual comparisons of different part features in PTE

As illustrated in Fig. 10, the matching accuracies of the baseline method are the worst because the correct samples have extremely similar global appearances to the incorrect samples. However, with the utilization of SFE and PTE, the matching accuracy is significantly improved. Our SFE module and PTE can extract the multi-scale and multi-part features from global appearances. They are useful in improving the ability of our method to distinguish similar samples. The retrieval results further validate the effectiveness of our proposed modules.

t-SNE visualization As shown in Fig. 11, we visualize the feature distributions of the baseline method and our CMT using t-SNE [59]. We randomly select 18 persons from the MSMT17 dataset, and 50 images of each person. Different colors represent different identities. From Fig. 11(a), it can be observed that the feature distributions with the same identity are relatively scattered. There are some misclassified samples. However, with our CMT, features of the same identity are more clustered and features of different identities are relatively separated. In addition, there are few misclassified samples compared with the baseline method. The t-SNE visualizations show that our method indeed helps the method learn a more discriminative embedding space, which further confirms our superiority to achieve robust person Re-ID.

5 Conclusion

In this paper, we integrate the advantages of CNNs and transformers and propose a novel learning framework named CMT for image-based person Re-ID. First, we propose a SFE module to extract the multi-scale features at different levels of the CNN backbone. Furthermore, we propose a PTE to generate and mine local diverse part features with global guidance. Experimental results on four public Re-ID benchmarks demonstrate that our method performs favorably against most state-of-the-art methods. In the future, we will reduce the computational complexity and improve the efficiency of our part-aware transformers.

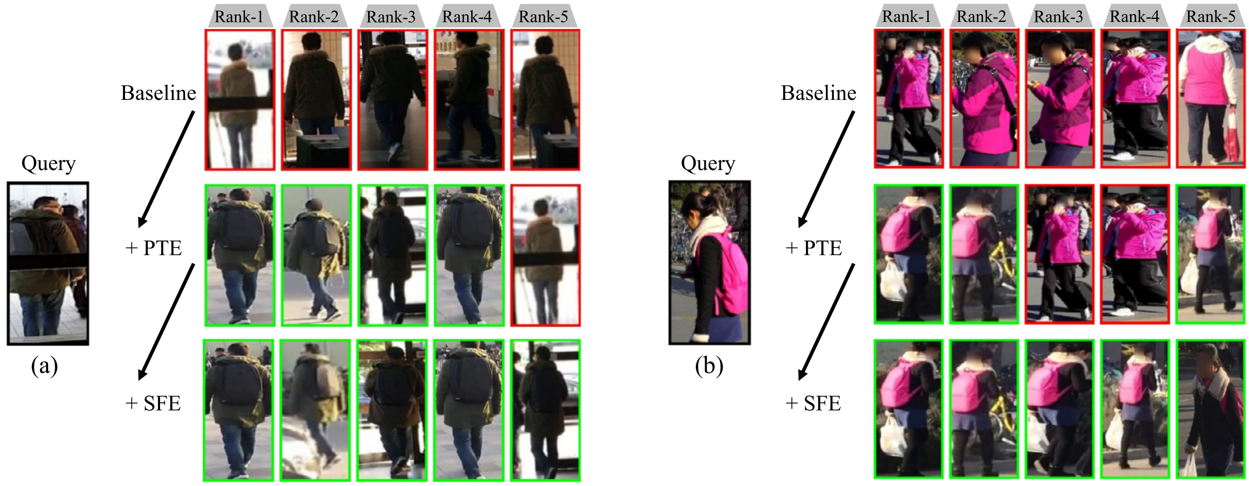

Figure 10 Visualization of the retrieved results on the MSMT17 dataset. The top-5 retrieved images are presented. The true matches are annotated by green boxes and the wrong matches are annotated by red boxes

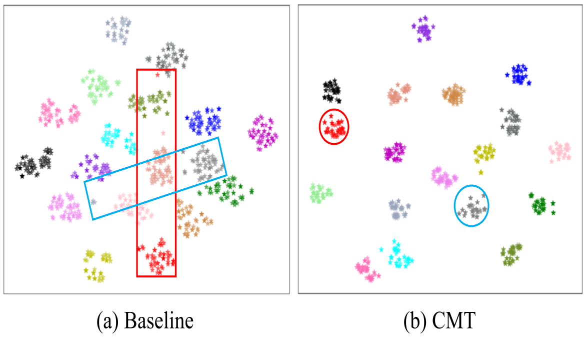

Figure 11 Comparison of feature distributions in Baseline and our CMT by using t-SNE. Different colors represent different identities

Acknowledgements
We thank the editor and all the reviewers for their constructive feedback.

Funding
This work was supported in part by the National Natural Science Foundation of China (NSFC) (No. 62101092) and the Fundamental Research Funds for the Central Universities (No. DUT20RC(3)083).

Abbreviations
CMC, cumulative matching characteristics; CMT, convolutional multi-level transformer; CNNs, convolutional neural networks; DSL, deeply-supervised learning; FFN, feed forward network; GAP, global average pooling; mAP, mean average precision; MHSA, multi-head self attention; PTE, part-aware transformer encoder; Re-ID, re-identification; SFE, scale-aware feature enhancement.

Availability of data and materials
The datasets analyzed during the current study have been publicly released and are available from the corresponding author upon reasonable request.

Code availability
The code is released at https://github.com/AI-Zhpp/CMT.

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by PY, XL and PZ. The first draft of the manuscript was written by PY and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Author details
[1]School of Artificial Intelligence, Dalian University of Technology, Dalian, 116024, China. [2]School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116024, China.

Received: 17 April 2023 Revised: 7 September 2023
Accepted: 11 September 2023 Published online: 13 October 2023

References

1. Loy, C. C., Xiang, T., & Gong, S. (2009). Multi-camera activity correlation analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1988–1995). Piscataway: IEEE.
2. Wang, X. (2013). Intelligent multi-camera video surveillance: a review. *Pattern Recognition Letters*, *34*, 3–19.
3. Zheng, L., Yang, Y., & Tian, Q. (2017). Sift meets CNN: a decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(5), 1224–1244.
4. Zheng, L., Yang, Y., & Hauptmann, A. G. (2016). Person re-identification: past, present and future. Preprint. arXiv:1610.02984.
5. Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2360–2367). Piscataway: IEEE.
6. Liu, C., Gong, S., Loy, C. C., & Lin, X. (2012). Person re-identification: what features are important? In A. Fusiello, V. Murino, & R. Cucchiara (Eds.), *Proceedings of the 12th European conference on computer vision* (pp. 391–401). Cham: Springer.
7. Shi, Z., Hospedales, T. M., & Xiang, T. (2015). Transferring a semantic representation for person re-identification and search. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4184–4193). Piscataway: IEEE.
8. Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: person retrieval with refined part pooling (and a strong convolutional

baseline). In F. Manhardt, W. Kehl, N. Navab, et al. (Eds.), *Proceedings of the 15th European conference on computer vision* (pp. 480–496). Cham: Springer.

9. Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). Omni-scale feature learning for person re-identification. In *2019 IEEE international conference on computer vision* (pp. 3702–3712). Piscataway: IEEE.
10. Wang, G., Yuan, Y., Chen, X., Li, J., & Zhou, X. (2018). Learning discriminative features with multiple granularities for person re-identification. In S. Boll, K. M. Lee, J. Luo, et al. (Eds.), *Proceedings of the 26th ACM international conference on multimedia* (pp. 274–282). New York: ACM.
11. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., et al. (2019). ABD-net: attentive but diverse person re-identification. In *2019 IEEE international conference on computer vision* (pp. 8351–8361). Piscataway: IEEE.
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Preprint. arXiv:1706.03762.
13. Zhu, K., Guo, H., Zhang, S., Wang, Y., Huang, G., Qiao, H., et al. (2021). Aaformer: auto-aligned transformer for person re-identification. Preprint. arXiv:2104.00921.
14. He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). TransReID: transformer-based object re-identification. In *2021 IEEE international conference on computer vision* (pp. 15013–15022). Piscataway: IEEE.
15. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., et al. (2019). Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8514–8522). Piscataway: IEEE.
16. Yang, Q., Yu, H.-X., Wu, A., & Zheng, W.-S. (2019). Patch-based discriminative feature learning for unsupervised person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3633–3642). Piscataway: IEEE.
17. Cho, Y., Kim, W. J., Hong, S., & Yoon, S. E. (2022). Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7308–7318). Piscataway: IEEE.
18. Chen, B., Deng, W., & Hu, J. (2019). In *2019 IEEE international conference on computer vision* (pp. 371–381). Piscataway: IEEE.
19. Rao, Y., Chen, G., Lu, J., & Zhou, J. (2021). Counterfactual attention learning for fine-grained visual categorization and re-identification. In *2021 IEEE international conference on computer vision* (pp. 1025–1034). Piscataway: IEEE.
20. Chen, G., Gu, T., Lu, J., Bao, J.-A., & Zhou, J. (2021). Person re-identification via attention pyramid. *IEEE Transactions on Image Processing*, *30*, 7663–7676.
21. Zhang, Z., Lan, C., Zeng, W., Jin, X., & Chen, Z. (2020). Relation-aware global attention for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3186–3195). Piscataway: IEEE.
22. Li, W., Zhu, X., & Gong, S. (2018). Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2285–2294). Piscataway: IEEE.
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 10th international conference on learning representations* (pp. 1–13). Retrieved August 25, 2023, from https://openreview.net/pdf?id=YicbFdNTTy.
24. Lai, S., Chai, Z., & Wei, X. (2021). Transformer meets part model: adaptive part division for person re-identification. In *2019 IEEE international conference on computer vision* (pp. 4150–4157). Piscataway: IEEE.
25. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y. D., & Wu, F. (2021). Diverse part discovery: occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2898–2907). Piscataway: IEEE.
26. Liao, S., & Shao, L. (2021). Transmatcher: deep image matching through transformers for generalizable person re-identification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, et al. (Eds.), *Advances in neural information processing systems 34* (pp. 1992–2003). Red Hook: Curran Associates.
27. Wang, G., Chen, X., Gao, J., Zhou, X., & Ge, S. (2021). Self-guided body part alignment with relation transformers for occluded person re-identification. *IEEE Signal Processing Letters*, *28*, 1155–1159.
28. Chen, X., Xu, J., Xu, J., & Gao, S. (2021). OH-Former: omni-relational high-order transformer for person re-identification. Preprint. arXiv:2109.11159.
29. Ma, Z., Zhao, Y., & Li, J. (2021). Pose-guided inter-and intra-part relational transformer for occluded person re-identification. In H. T. Shen, Y. Zhuang, J. R. Smith, et al. (Eds.), *Proceedings of the 29th ACM international conference on multimedia* (pp. 1487–1496). New York: ACM.
30. Liu, X., Zhang, P., Yu, C., Lu, H., Qian, X., & Yang, X. (2021). A video is worth three views: trigeminal transformers for video-based person re-identification. Preprint. arXiv:2104.01745.
31. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Piscataway: IEEE.
32. Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. Preprint. arXiv:1511.07122.
33. Zhang, G., Zhang, P., Qi, J., & Lu, H. (2021). HAT: hierarchical aggregation transformers for person re-identification. In H. T. Shen, Y. Zhuang, J. R. Smith, et al. (Eds.), *Proceedings of the 29th ACM international conference on multimedia* (pp. 516–525). New York: ACM.
34. Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Amulet: aggregating multi-level convolutional features for salient object detection. In *2017 IEEE international conference on computer vision* (pp. 202–211). Piscataway: IEEE.
35. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826). Piscataway: IEEE.
36. Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. Preprint. arXiv:1703.07737.
37. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. D., & Tian, Q. (2015). Scalable person re-identification: a benchmark. In *2015 IEEE international conference on computer vision* (pp. 1116–1124). Piscataway: IEEE.
38. Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *2017 IEEE international conference on computer vision* (pp. 3754–3762). Piscataway: IEEE.
39. Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). 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). Piscataway: IEEE.
40. Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 79–88). Piscataway: IEEE.
41. Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the 34th AAAI conference on artificial intelligence* (pp. 13001–13008). Palo Alto: AAAI Press.
42. Kingma, D. P., & Ba, J. (2014). Adam: a method for stochastic optimization. Preprint. arXiv:1412.6980.
43. Chen, X., Fu, C., Zhao, Y., Zheng, F., Song, J., Ji, R., et al. (2020). Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3300–3310). Piscataway: IEEE.
44. Si, J., Zhang, H., Li, C.-G., Kuen, J., Kong, X., Kot, A. C., et al. (2018). Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5363–5372). Piscataway: IEEE.
45. Wang, C., Zhang, Q., Huang, C., Liu, W., & Wang, X. (2018). Mancs: a multi-task attentional network with curriculum sampling for person re-identification. In F. Manhardt, W. Kehl, N. Navab, et al. (Eds.), *Proceedings of the 15th European conference on computer vision* (pp. 365–381). Cham: Springer.
46. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019). Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9317–9326). Piscataway: IEEE.
47. Luo, H., Gu, Y., Liao, X., Lai, S., & Jiang, W. (2019). Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1487–1495). Piscataway: IEEE.
48. Kalayeh, M. M., Basaran, E., Gökmen, M., Kamasak, M. E., & Shah, M. (2018). Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1062–1071). Piscataway: IEEE.
49. Tay, C.-P., Roy, S., & Yap, K.-H. (2019). AANet: attribute attention network for person re-identifications. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7134–7143). Piscataway: IEEE.

50. Zheng, M., Karanam, S., Wu, Z., & Radke, R. J. (2019). Re-identification with consistent attentive Siamese networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5735–5744). Piscataway: IEEE.
51. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., & Zhang, S. (2019). Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1389–1398). Piscataway: IEEE.
52. Fang, P., Zhou, J., Roy, S. K., Petersson, L., & Harandi, M. (2019). Bilinear attention networks for person retrieval. In *2019 IEEE international conference on computer vision* (pp. 8030–8039). Piscataway: IEEE.
53. Dai, Z., Chen, M., Gu, X., Zhu, S., & Tan, P. (2019). Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3691–3701). Piscataway: IEEE.
54. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., & Kautz, J. (2019). Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2138–2147). Piscataway: IEEE.
55. Jin, X., Lan, C., Zeng, W., Chen, Z., & Zhang, L. (2020). Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3143–3152). Piscataway: IEEE.
56. Zhu, K., Guo, H., Liu, Z., Tang, M., & Wang, J. (2020). Identity-guided human semantic parsing for person re-identification. Preprint. arXiv:2007.13467.
57. Xu, B., He, L., Liao, X., Liu, W., Sun, Z., & Mei, T. (2020). Black Re-ID: a head-shoulder descriptor for the challenging problem of person re-identification. In C. W. Chen, R. Cucchiara, X.-S. Hua, et al. (Eds.), *Proceedings of the 28th ACM international conference on multimedia* (pp. 673–681). New York: ACM.
58. Li, H., Wu, G., & Zheng, W.-S. (2021). Combined depth space based architecture search for person re-identification. Preprint. arXiv:2104.04163.
59. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.