Transformer Meets Part Model: Adaptive Part Division for Person Re-Identification

Shenqi Lai
Meituan
laishenqi@meituan.com

Zhenhua Chai
Meituan
chaizhenhua@meituan.com

Xiaolin Wei
Meituan
weixiaolin02@meituan.com

Abstract

Part model is one of the key factors to high performance person re-identification (ReID) task. In recent studies, there are mainly two streams for part model. The first one is to divide a person image into several fixed parts to obtain their local information, but it may cause performance degradation in case of misalignment. The other one is to explore external resources like pose estimation or human parsing to locate local parts, but it costs extra storage and computation. Inspired by recent successful transformers on spatial similarity modeling, we propose a novel Adaptive Part Division (APD) model to better extract local features. More specifically, APD mainly consists of two crucial modules: a Transformer-based Part Merge (TPM) module and a Part Mask Generation (PMG) module. In particular, TPM first adaptively assigns the patch tokens of the same semantic object to the identical part. Then, PMG takes these identical parts together and generates several non-overlapping masks for robust part division. We have conducted extensive evaluations on four popular benchmarks, i.e. Market-1501, CUHK03, DukeMTMC-ReID and MSMT17, and the experimental results show that our proposed method achieves the state-of-the-art performance.

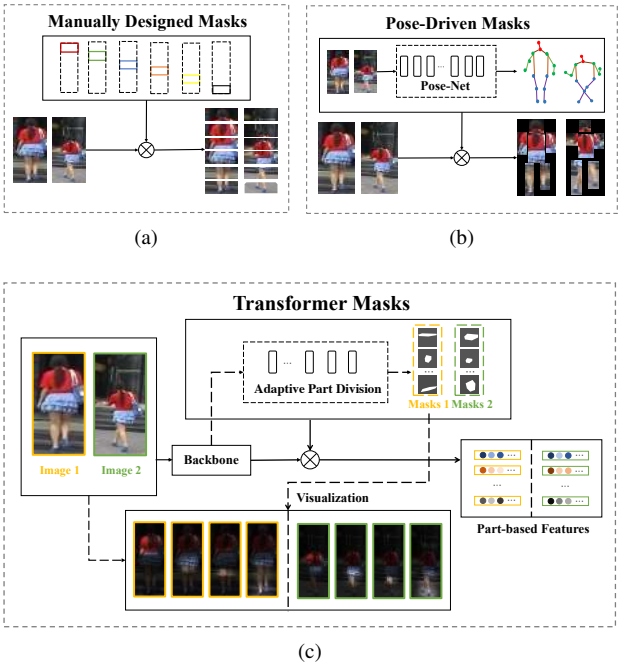

Figure 1: Three different methods for local feature extraction. a) Manually Designed Masks. b) Pose-Driven Masks. c) Our Transformer Masks.

1. Introduction

Person re-identification (ReID) [40] has been an active research topic in computer vision and machine learning techniques, which is an essential component in some important applications such as intelligent visual surveillance (IVS) and driverless car. The aim of ReID is to identify whether an individual has already been observed over a camera in a network and the task is very challenging due to the complicated environment (e.g. illumination, pose and partial occlusion) and even different camera views.

Thanks to the recent development of the Convolutional Neural Network (CNN) [14], the part based representations [26, 29] for ReID which can capture finer information exhibit competitive performance and some typical ReID applications of this kind have been deployed in practice. However, the performance of ReID will probably be affected by the preceding step pedestrian detection, the study on improving robustness is necessary.

Generally speaking, for ReID task the person will first be cropped from the image with the bounding box provided by the pedestrian detecter, and the bounding boxes are assumed to be precise and highly-aligned. Despite the recent great progress, detected results can not always be promising and there might be some biases in scale and shift, which will lead to misalignment.

For the part-based model in Figure 1 (a), images will be

divided by fixed masks. Since different pedestrian images share the same masks, parts will be divided by mistake if there is misalignment. In this way, the part-level features which contain foot information will be compared with features derived from thigh in the other image, and this will lead to failure matching. A direct and effective way to solve this problem is to use external resources like pose estimation [23] or human parsing [13], which is shown in Figure 1 (b). Pedestrian landmarks predicted by a pose estimation network are used to help extract local features by pose-aware pooling. However, the external network will increase the training and inference time significantly, and in some cases the pose estimation methods are even more time-consuming than person ReID itself.

In this paper, we propose a novel Adaptive Part Division model to address above issues, which is shown in Figure 1 (c). Instead of setting the part masks manually, our proposed APD can produce the masks adaptively according to the input image. Firstly, the mask values are computed adaptively in Transformer-based Part Merge module according to the input image, which is expected to be robust to scale and shift variations. Secondly, in order to make the extracted features more discriminative, Part Mask Generation module also introduce competitive mechanism to assure the diversity among different masks. Thirdly, the transformer masks are derived from the existing feature extraction backbone, in this way it will not bring too much extra computation. Our contributions can be summarized as follows:

1) We have proposed a novel Adaptive Part Division (APD) model for part-based ReID feature extraction. Unlike traditional methods which use manually designed masks, the proposed APD containing TPM and PMG still has the chance to capture the corresponding areas of two different samples even when they meet scale and shift misalignments.

2) APD is generally applicable and model-agnostic. It can easily be applied to most of the existing popular part-based architectures, such as PCB [26] and MGN [29].

3) Extensive experimental results demonstrate the superiority of the proposed method over a wide range of the state-of-the-art ReID models on four large benchmarks, i.e. Market-1501 [39], DukeMTMC-ReID [21, 42], CUHK03-NP [16, 43] and MSMT17 [31].

The rest part of this paper is organized as follows. Section 2 will introduce the related works in person ReID. Section 3 is about the design details of APD. The extensive experiments have been conducted and the results will be discussed in Section 4. Finally in Section 5 the conclusion will be drawn.

2. Related Work

Feature extraction is the key step for person ReID, and the aim is to obtain discriminative features. Inspired by the success of CNN on image classification, deep global-based methods exhibit some new insights in this direction. A simple way [3] to extract global features is applying CNN to pedestrian images, and the relations among different samples can be further explored to enhance the discriminability. BagOfTricks [20] is also a typical work of this kind, which improves the performance by combining a collection of training tricks. However, in some complex scenarios the performance will probably be affected from large variations like pose, occlusion and background clutters.

2.1. Part model for ReID

Part-based methods, which are designed to focus on local regions and capture fine-grained cues, are expected to be more effective and robust to these challenges. Methods such as [16, 17, 26, 29, 38, 9, 15, 28] generate the final representations with specific predefined semantic parts, and the fused features usually perform better than the original global counterparts. However, these methods still have some shortcomings. When used in the real scenes, the predefined rigid body parts will not be robust to large pose deformations and complex view variations.

To solve this problem, an effective method is using external resources. SPReID [13] employs human semantic parsing network to harness local visual cues. Semantic segmentation results not only help to align each human part, but also reduce the disturbance of complex background. In PDC model [23], pose transformation network is used to do affine transform for cropped part regions. The learned local representations will hence focus on the transformed regions. PABR [24] adopts a simple but effective solution. Features from OpenPose [1] and GoogleNet [27] are aggregated by bilinear pooling [19] to get final part-aligned features.

Methods with external resources mentioned above achieve great success, but all of them can not be avoided to import extra networks int both training and testing stage. In this paper we propose an end-to-end learning framework, on which the local feature extraction of different salient body parts can be implemented with a novel Adaptive Part Division model. Different from the existing methods, our proposal will not involve any models from extra resources.

2.2. Transformer and Attention model for ReID

Recently, attention mechanism is a popular architectural in neural networks, which improves the performance in both natural language processing and computer vision. In recent successful Vision Transformer, it also plays a key role. In person ReID, transformer and attention model is used to deal with misalignment problem. HA-CNN [18] combines

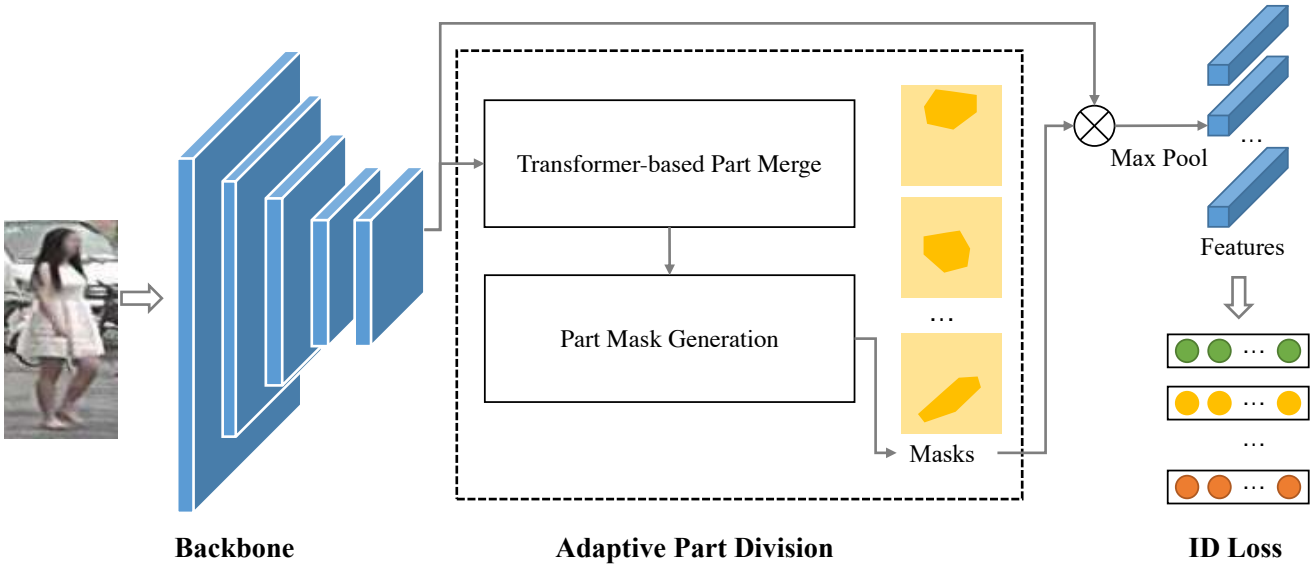

Figure 2: Overview of our Framework.

both soft attention and hard attention to optimise misaligned person images. IANet [12] proposes two attention mechanism named SIA (Spatial Interaction-and-Aggregation) and CIA (Channel Interaction-and-Aggregation) to model the large variations in person pose and scale. ABD-Net [4] adopts similar methods and proposes orthogonality regularization to both hidden features and weights. TransReID [11] is a pure transformer-based ReID framework, which encodes one image as a sequence of patches and build a transformer-based strong baseline. AAformer [45] adopts the "part tokens" to learn the part representations and integrates the part alignment into the self-attention. HAT [30] aggregate multi-scale features to better fuse semantic and detail information for image-based person ReID. Instead of producing attention feature maps like previous researches, we are more interested in generating discriminative masks and features at same time, which could produce better local features.

3. Proposed Method

In this section, we will introduce the technical details of the proposed Adaptive Part Division (APD) model and the flowchart can be found in Figure 2 and Figure 3. There are mainly two innovations on APD and they are fused into a single framework. The first part is Transformer-based Part Merge (TPM) module and will be described in Section 3.1. It is expected to enhance the representation and make it robust to scale and shift variations. The second part is Part Mask Generation (PMG) module, which is designed to avoid generate highly similar masks. Different masks are supposed to focus on different key regions. The details will be introduced in Section 3.2.

3.1. Transformer-based Part Merge Module

The main disadvantage for manual division is that both mask size and mask position are fixed, and as we analyze before the ReID performance will probably be affected by detection results. During training, transformer masks are expected to locate serval informative and salient regions according to the input images meanwhile the feature map will be refined and updated iteratively according to structural information. In order to well generate the discriminative transformer masks, it is important to provide discriminative feature maps. We adopt Transformer-based Part Merge, which could model relations between all local patches on feature maps, to produce distinctive part regions.

The structure of TPM is illustrated in Figure 3, which could not only capture local part information but also can explore the long-range dependencies. We follow Vision Transformer (ViT) [6] to construct the main module. Specially, class token and position embeddings are not adopted. The former is designed to produce the embedding for classification, but in this task, we only aim to cluster similar patch embeddings to one part. The latter is used to retain positional information to the vector sequence in element-wise, however, our enhanced feature maps will finally generate masks in PMG, and positional information has no benefit to this.

Given a feature map $x \in \mathbb{R}^{H\times W\times C}$ with resolution (H,W) and channel C as input, we reshape it into a sequence of flattened 2D patches $x \in \mathbb{R}^{N\times C}$ to fit the Transformer architecture, and $N = (H \times W)$ is the length of feature patch sequence. Then, We map the patches to vec-

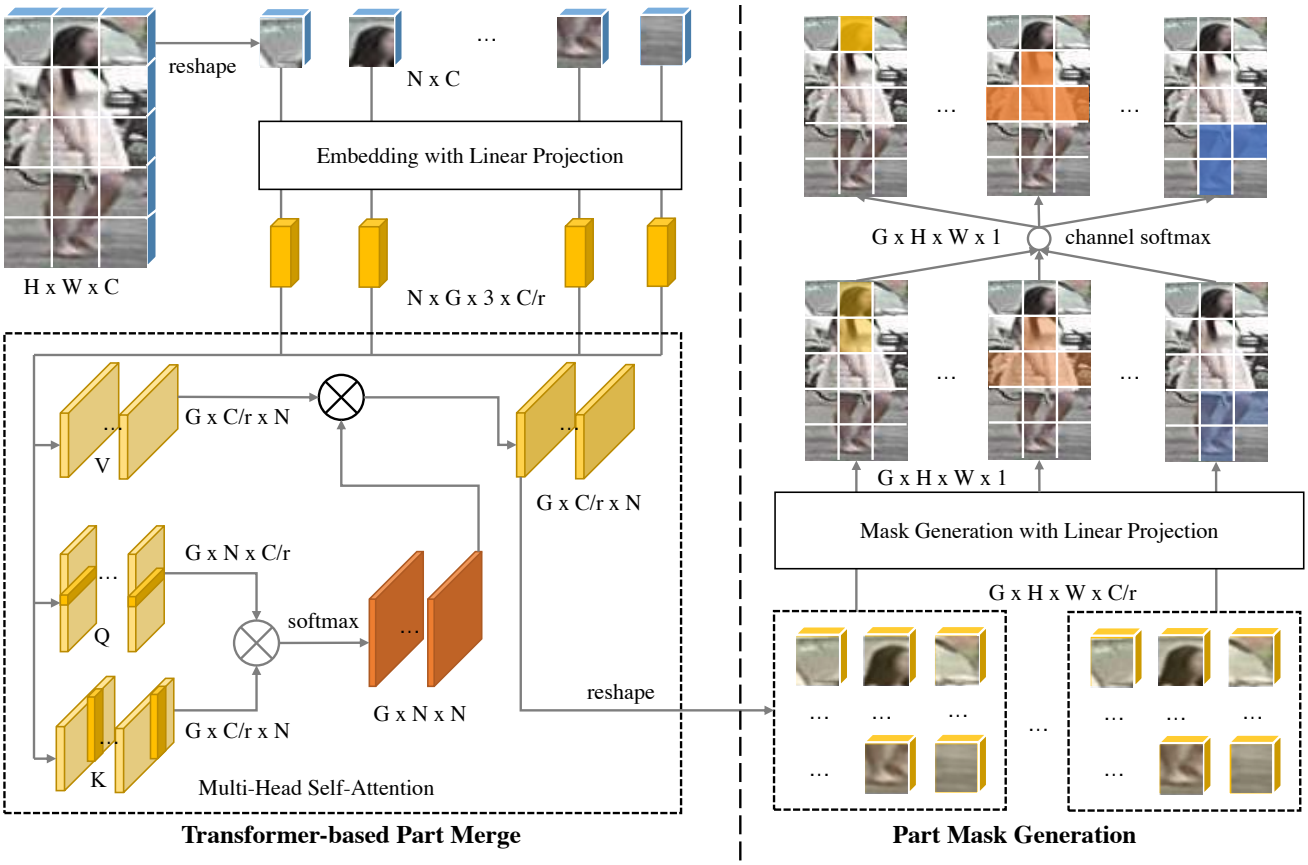

Figure 3: Details of TPM and PMG.

tors of $N \times G \times 3 \times C/r$ dimensions with linear projection and modify them as the input of transformer encoder. The linear projection is a 1×1 convolution layer to reduce dimensions. G means attention heads in Multi-head Self-Attention (MSA), and r is the reduction ratio. The outcome vector sequence $Z \in \mathbb{R}^{N \times C/r}$ is fed to a MSA encoder, Z means the query, key or value in one attention head. Unlike the standard Transformer encoder, we remove Multi-Layer Perception (MLP) blocks, because these blocks will impact the diversity of masks in PMG.

The self-attention mechanism is based on three types of vectors, query, key and value. Query vector in a sequence ($Q \in \mathbb{R}^{N \times C/r}$), we match it against a set of key vectors ($K \in \mathbb{R}^{N \times C/r}$) using inner products. These inner products are then scaled and normalized with a softmax function to obtain the score matrix ($S \in \mathbb{R}^{N \times N}$). The output of the self-attention for this query is the weighted sum of a set of N value vectors ($V \in \mathbb{R}^{N \times C/r}$). For all the queries in the sequence, the output matrix of self-attention can be obtained by:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{D}}V), \quad (1)$$

where the Softmax function is applied over each row of the input matrix and the $\sqrt{D}$ term provides appropriate normalization. Query, key and value matrices are all computed from the outcome vector sequence by vector splitting, we could also regard them as three different linear projection: $Q = ZW_Q, K = ZW_K, V = ZW_V$.

Finally, the MSA layer is defined by considering G attention "heads", which means G self-attention functions are applied to the input in parallel. Each head provides a sequence of size $N \times C/r$, we do not any rearrangments, because they are fed into PMG module, and each head will produce a transformer mask.

3.2. Part Mask Generation Module

Though we generate different mask from different attention head, it is still hard to to guarantee that all masks could highlight distinct local regions. In some cases, these masks may be homogeneous. Thus, it is important to guide each mask to focus on different regions to reduce redundancy.

In other words, reducing the overlap area among different regions is key to this problem. To achieve this goal, we introduce competitive mechanism to solve the problem. Same positional features on different masks should only

have one maximum response value, which means that only one mask could have response to one position.

As shown in Figure 3, the outputs of TPM is reshaped to $H \times W \times C/r$ on each head independently. Then, we do 1×1 convolution to all reshaped feature maps to get initial masks $M_0 \in \mathbb{R}^{G \times H \times W \times 1}$. Specially, this linear projection is adopted separately, to avoid generating homogeneous masks. Then, a softmax function is used to establish competitive mechanism on channel dimension. It could further adjust initial masks, by limiting same position response on different masks. We obtain competitive masks $M_1 \in \mathbb{R}^{G \times H \times W \times 1}$, and regard them as diverse salient regions for a person. They are adaptive to the input image, and have good diversity to capture important regions.

4. Experiments

4.1. Datasets and Settings

Three popular datasets are used to evaluate the performance of our proposed method. Market-1501 dataset [39] is one of the most frequently used datasets, which contains 1,501 labeled persons of 6 camera viewpoints. There are 12,936 images with 751 identities for training. The rest with 750 identities is in the testing set which contains 3,368 query images and 19,732 gallery images. DukeMTMC-ReID [21, 42] is a subset of the DukeMTMC dataset. It is also one of the most challenging ReID datasets up to now, which contains 1,404 identities captured by 8 cameras in realistic conditions in winter. There are 16,522 training images, 2,228 query images and 17,661 gallery images. CUHK03 [16] dataset contains 13,164 images of 1,467 identities. Each identity is observed by 2 cameras. Different from the former two datasets, CUHK03 offers both hand-labeled and DPM-detected [8] bounding boxes, all of them are used in our evaluation. Considering the time-consuming evaluation, we adopt the new training/testing protocol proposed in [43], instead of adopting 20 random splits in original paper. MSMT17 [31] is the largest person ReID dataset. It have 126,441 images with 4,101 identities on 15 cameras. It is worthy mentioning that we strictly follow the commonly used protocol for Market-1501 and DukeMTMC-ReID datasets. In all experiments, we evaluate the results with only the single-query setting and the multi-query setting is not adopted. In addition, we do not apply re-ranking [43] algorithm in order to have a fair comparison with all methods above.

4.2. Implementation Details

We conduct all the experiments using Pytorch with 4 NVIDIA TESLA V100 GPUs. For data augmentation, we use horizontal flipping, random erasing and normalization. In addition, all the images are resized to resolution 384×128, and after padding 10px with value 0 the random region crop sized of 384×128 will be used as the final output. We use the pretrained ResNet-50 and ResNet-101 [10] as the backbone network. The only difference is we set the stride to 1 in the last block. All networks are trained with SGD. Besides, batch size, weight decay and momentum will be set to 64, 5e-4 and 0.9 respectively. We apply linear warm-up strategy for the first 5 epochs, and the learning rate will be increased from 1e-3 to 1e-1. We adopt Cosine learning rate during training and the total number of epochs is set to 90, which will take 60 minutes for ResNet-50 and 110 minutes for ResNet-101. For the hyper parameters in our method, the reduction ratio in attention mask is set to 64. The attention head G is set to 6, and the reduction ratio r is set to 32. We use max pooling instead of avg pooling to extract the features because MGN [29] finds the former has a better result.

Method	Backbone	Rank-1	mAP
AACN [33]	GoogleNet	85.9	66.9
PL-Net [36]	GoogLeNet	88.2	69.3
SPReID [13]	InceptionV3	92.5	81.3
PCB [26]	ResNet-50	93.8	81.6
PABR [24]	GoogleNet	91.7	79.6
MGN [29]	ResNet-50	95.7	86.9
ADReID [34]	ResNet	95.0	86.5
VPM [25]	ResNet-50	93.0	80.8
BagOfTricks [20]	ResNet-50	94.5	85.9
HPM [9]	ResNet-50	94.2	82.7
PPS [22]	ResNet-50	94.3	85.3
IANet [12]	ResNet-50	94.4	83.1
DGNet [41]	ResNet-50	94.8	86.0
DSAReID [37]	ResNet-50	95.7	87.6
Pyramid [38]	ResNet-101	95.7	88.2
CAMA [35]	ResNet-50	94.7	84.5
BDB [5]	ResNet-50	94.2	84.3
OSNet [44]	ResNet-50	94.8	84.9
MHN-6 [2]	ResNet-50	95.1	85.0
SONA [32]	ResNet-50	95.6	88.8
BAT-net [7]	GoogleNet-BN	95.1	87.4
ABD-Net [4]	ResNet-50	95.6	88.3
TransReID [11]	DeiT-B/16	94.9	88.1
AAformer [45]	ViT-B/16	95.4	87.7
HAT [30]	ResNet-50	95.6	89.5
Ours (PCB)	ResNet-50	95.5	87.5
Ours (PCB)	ResNet-101	95.6	88.6
Ours (MGN)	ResNet-50	95.8	89.1
Ours (MGN)	ResNet-101	**96.0**	**90.3**

Table 1: Comparison with the state-of-the-art methods on the Market-1501.

Method	Backbone	Rank-1	mAP
AACN [33]	GoogleNet	76.8	59.3
SPReID [13]	InceptionV3	84.0	69.8
PCB [26]	ResNet-50	83.3	69.2
PABR [24]	GoogleNet	84.4	69.3
MGN [29]	ResNet-50	88.7	78.4
ADReID [34]	ResNet	86.0	74.6
VPM [25]	ResNet-50	83.6	72.6
BagOfTricks [20]	ResNet-50	86.4	76.4
HPM [9]	ResNet-50	86.6	74.3
PPS [22]	ResNet-50	86.6	74.3
IANet [12]	ResNet-50	87.1	73.4
DGNet [41]	ResNet-50	86.6	74.8
DSAReID [37]	ResNet-50	86.2	74.3
Pyramid [38]	ResNet-101	89.0	79.0
CAMA [35]	ResNet-50	85.8	72.9
BDB [5]	ResNet-50	89.0	76.0
OSNet [44]	ResNet-50	86.6	74.8
MHN-6 [2]	ResNet-50	89.1	77.1
SONA [32]	ResNet-50	89.4	78.3
BAT-net [7]	GoogleNet-BN	87.7	77.3
ABD-Net [4]	ResNet-50	89.0	78.6
AAformer [45]	ViT-B/16	90.1	80.0
TransReID [11]	DeiT-B/16	90.2	81.3
HAT [30]	ResNet-50	90.4	81.4
Ours (PCB)	ResNet-50	87.1	74.2
Ours (PCB)	ResNet-101	88.1	75.4
Ours (MGN)	ResNet-50	90.7	81.1
Ours (MGN)	ResNet-101	**91.3**	**82.1**

Table 2: Comparison with the state-of-the-art methods on the DukeMTMC-ReID.

Method	Backbone	Labeled		Detected	
		Rank-1	mAP	Rank-1	mAP
PCB [26]	ResNet-50	-	-	63.7	57.5
MGN [29]	ResNet-50	68.2	67.4	66.8	66.0
HPM [9]	ResNet-50	-	-	63.9	57.5
PPS [22]	ResNet-50	75.6	72.7	73.7	70.6
DSAReID [37]	ResNet-50	78.9	75.2	78.2	73.1
Pyramid [38]	ResNet-101	78.9	76.9	78.9	74.8
CAMA [35]	ResNet-50	70.1	66.5	66.6	64.2
BDB [5]	ResNet-50	-	-	76.4	73.5
BAT-net [7]	GoogleNet-BN	78.6	76.1	76.2	73.2
OSNet [44]	ResNet-50	-	-	72.3	67.8
MHN-6 [2]	ResNet-50	77.2	72.4	71.7	65.4
AAformer [45]	ViT-B/16	79.9	77.8	77.6	74.8
HAT [30]	ResNet-50	**82.6**	80.0	79.1	75.5
Ours (PCB)	ResNet-50	77.0	73.8	74.6	70.6
Ours (PCB)	ResNet-101	78.7	75.9	77.0	73.4
Ours (MGN)	ResNet-50	79.9	77.2	78.1	75.3
Ours (MGN)	ResNet-101	82.1	**80.6**	**79.2**	**75.7**

Table 3: Comparison with the state-of-the-art methods on the CUHK03-NP

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

In order to show the effectiveness of our proposed method, we have compared it with the state-of-the-art methods on Market-1501, DukeMTMC-ReID, CUHK03-NP and MSMT17. We adopt the same architecture with PCB as backbone, and the experimental results show that our proposed method is comparable with the state-of-the-art methods. In addition, if we use more complex structure (e.g. MGN [29], a multi-branch architecture), our method can be further boosted on all four datasets. Replacing ResNet-50 backbone with ResNet-101 leads to significant improvements.

Method	Backbone	Rank-1	mAP
IANet [12]	ResNet-50	75.5	46.8
PCB [26]	ResNet-50	68.2	40.4
BAT-net [7]	GoogleNet-BN	79.5	56.8
ABD-Net [4]	ResNet-50	82.3	60.8
AAformer [45]	ViT-B/16	**83.1**	62.6
HAT [45]	ResNet-50	82.3	61.2
Ours (PCB)	ResNet-50	79.8	57.1
Ours (PCB)	ResNet-101	80.3	59.9
Ours (MGN)	ResNet-50	82.4	61.2
Ours (MGN)	ResNet-101	82.9	**62.7**

Table 4: Comparison with the state-of-the-art methods on the MSMT17.

Method	Rank-1	Rank-5	Rank-10	mAP
PCB	94.8	98.2	98.8	85.4
+ PMG (sigmoid)	95.0	98.3	98.9	86.6
+ PMG (w/o softmax)	95.1	98.3	98.8	86.4
+ PMG (softmax)	95.3	98.3	98.9	86.5
+ TPM (w/o heads) + PMG	95.4	98.3	98.0	87.1
+ TPM + PMG (w/o heads)	95.3	98.4	98.9	87.2
+ TPM + PMG	**95.5**	**98.6**	**99.2**	**87.5**
+ TPM + PMG + CLS	95.5	98.6	99.2	87.4
+ TPM + PMG + ABS PE	95.4	98.3	98.9	87.1
+ TPM + PMG + REL PE	95.3	98.3	99.1	87.4

Table 5: Details of the module settings.

More specifically, HAT is also based on ResNet-50, and the performance is better than our method on PCB with ResNet-50. However, the result with MGN for our method is comparable with HAT, the Rank-1 is better and the mAP is worse. Further more, HAT adopts four standard transformer modules on four stages. According to our calculations, the extra FLOPs is 1460M, which is almost 40% of ResNet-50. While we only use one transformer module, and the FLOPs is 481M, which is much less than HAT. More details can be found in Table 1, Table 2, Table 3 and Table 4.

4.4. Ablation Study

To verify the effectiveness of each component, we have designed extensive ablation studies on Market-1501, including mask strategies and reduction ratio. In addition, all the rest settings are kept the same as in Section 4.2.

Mask Strategies. As shown in Table 5, TPM gives us significant improvement over basic PCB, and using channel softmax to form a competitive mechanism further improve the performance. Adding TPM with multi heads could en-

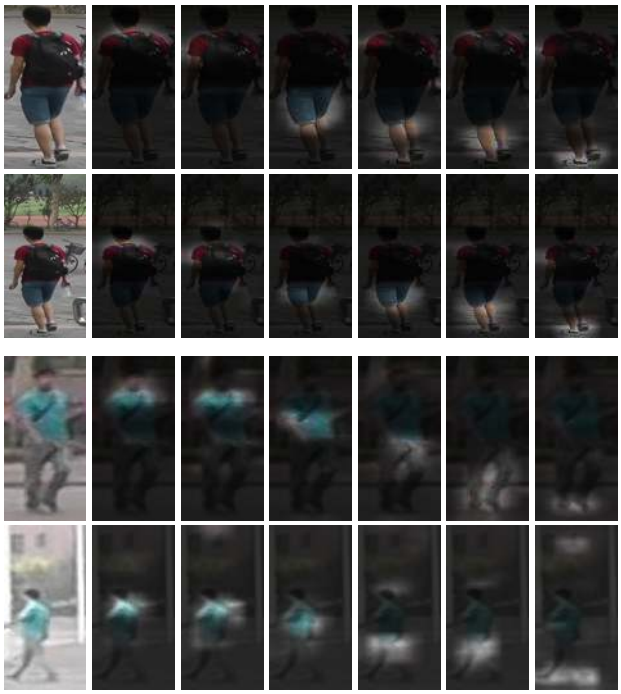

Figure 4: Visualization of some examples. Even though there are scale and shift misalignments, our proposed APD can still capture the corresponding areas.

Reduction Ratio	Rank-1	Rank-5	Rank-10	mAP
1	95.0	98.2	98.9	86.8
2	95.0	98.3	98.9	86.7
4	95.1	98.1	98.9	87.0
8	95.1	98.1	98.9	87.0
16	95.0	98.3	98.9	87.1
32	**95.5**	**98.6**	**99.2**	**87.5**
64	95.4	98.4	99.0	87.4
128	95.2	98.2	98.8	87.0
256	95.1	98.4	98.9	87.2

Table 6: Performance for different reduction ratios in TPM.

hance discriminability of features and also increase diversity of masks. It is also beneficial to the result. We also add class token and we find it has no use fot the result. For position embedding, both absolute position and relative position are harmful to the performance. The details on the visualization of the masks can be found in Fig. 4.

Reduction Ratio in TPM. As shown in Table 6, different reduction ratios can change performance significantly. The best result achieves 95.5% for Rank-1 and 87.5% for mAP, while the worst results for Rank-1 and mAP are only 95.0% and 86.7% respectively. We find that both large number or small number of channels in TPM leads to poor performance, the reason could be that large number of channels may introduce too much noises in similarity calculation, and small number is not enough to represent a patch. Proper channel number helps to learn the better corelation among different local patches.

5. Conclusion

In this paper, we propose a novel method named Adaptive Part Division (APD) for person ReID, which contains a Transformer-based Part Merge (TPM) module and a Part Mask Generation (PMG) module. TPM introduces Transformer to enhance the discriminability of features and assigns the patch tokens of the same semantics to one identical part. PMG produces non-overlapping part marks adaptively, which increases robustness to scale and shift variations. Extensive experiments have been conducted on four ReID datasets, and the results show that our proposed method is comparable with the state-of-the-arts.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

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

[3] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 2016.

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

[5] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *ICCV*, 2019.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *ICCV*, 2019.

[8] Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[9] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019.

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

[11] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv:2102.04378*, 2021.

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

[13] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[15] Hui Li, Meng Yang, Zhihui Lai, Weishi Zheng, and Zitong Yu. Pedestrian re-identification based on tree branch network with local and global learning. In *ICME*, 2019.

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

[17] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017.

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

[19] Tsung-Yu Lin, Aruni Roy Chowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *ICCV*, 2015.

[20] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshop*, 2019.

[21] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop*, 2016.

[22] Yunhang Shen, Rongrong Ji, Xiaopeng Hong, Feng Zheng, Xiaowei Guo, Yongjian Wu, and Feiyue Huang. A part power set model for scale-free person retrieval. In *IJCAI*, 2019.

[23] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

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

[25] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, 2019.

[26] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In *ECCV*, 2018.

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[28] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *AAAI*, 2019.

[29] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018.

[30] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, 2021.

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

[32] Bryan (Ning) Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *ICCV*, 2019.

[33] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[34] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention driven person re-identification. *Pattern Recognition*, 2019.

[35] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, 2019.

[36] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 2019.

[37] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019.

[38] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, 2019.

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

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

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

[42] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[43] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[44] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, 2019.

[45] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *arXiv:2104.00921*, 2021.