This ACCV 2020 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2020
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

This is the Pre-Published Version.

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use (https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms), but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-030-69538-5_9.

Part-aware Attention Network for Person Re-Identification

Wangmeng Xiang[1 [0000−0002−4373−2610]], Jianqiang Huang[2 [0000−0001−5735−2910]], and Xian-Sheng Hua[2 [0000−0002−8232−5049]] Lei Zhang[1,2 [0000−0002−2078−4215]]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{cswxiang,cslzhang}@comp.polyu.edu.hk
[2] Artificial Intelligence Center, Alibaba DAMO Academy, Hangzhou, China
{jianqiang.hjq,xiansheng.hxs}@alibaba-inc.com

Abstract. Multi-level feature aggregation and part feature extraction are widely used to boost the performance of person re-identification (Re-ID). Most multi-level feature aggregation methods treat feature maps on different levels equally and use simple local operations for feature fusion, which neglects the long-distance connection among feature maps. On the other hand, the popular horizon pooling part based feature extraction methods may lead to feature misalignment. In this paper, we propose a novel Part-aware Attention Network (PAN) to connect part feature maps and middle-level features. Given a part feature map and a source feature map, PAN uses part features as queries to perform second-order information propagation from the source feature map. The attention is computed based on the compatibility of the source feature map with the part feature map. Specifically, PAN uses high-level part features of different human body parts to aggregate information from mid-level feature maps. As a part-aware feature aggregation method, PAN operates on all spatial positions of feature maps so that it can discover long-distance relations. Extensive experiments show that PAN achieves leading performance on Re-ID benchmarks Market1501, DukeMTMC, and CUHK03.

1 Introduction

Person re-identification (Re-ID) aims to recognize a person of interest from the gallery by using a query image or video clip. The state-of-the-art person Re-ID methods [1–6] usually employ a convolutional neural network (CNN) to extract the feature vector of a person, and use metric learning or classification loss to enforce the learned feature representations to be discriminative. How to design an efficient and effective network structure for feature extraction is critical for high-performance person Re-ID.

Recent CNN-based person Re-ID methods mainly fall into two territories. 1) The first is creating more effective metrics to fully explore the relations between features [7–12]. For example, Cheng *et al.* [13] designed a multi-channel part-based CNN model under the triplet framework. Chen *et al.* [12] applied a quadruplet loss with a margin based online hard negative mining. 2) The

second is designing efficient and effective architecture for feature extraction [1, 14–18], where off-the-shelf feature extractors and prior knowledge about the human body are employed. For instance, PCB [5] utilizes ResNet50 backbone and applies horizon strip pooling and refined part pooling to extract discriminative and mutual distinctive human part features. MGN [19] employs a multi-branch network and applies horizon strip pooling on each branch.

One popular direction for effective architecture design is fusing features from different layers of a network. Existing multi-scale and multi-level Re-ID methods [16, 20–23] typically use multi-branch or directly apply convolutional layers on mid-level features to generate multi-level feature maps, and fuse the feature maps by element-wise operation. However, combining different levels of features in this way has several problems. First, the spatial resolution of high-level features is reduced largely, and fine details may be lost during the down-sampling process, making the feature fusion less effective. Second, both residual and concatenation connections are local operations, neglecting the long-range relationships between different layers of feature maps. Last but not least, due to the lack of guidance, background, noise, and distractors will also be fused to the target feature maps and result in inferior performance.

In this paper, we present a Part-aware Attention Network (PAN) by using a Part-aware Attention Module (PAM) to bridge different layers' feature maps in a CNN. A part feature map $\mathbf{P}$ is first generated. Then given feature maps $\mathbf{X}$, PAM takes a feature vector in $\mathbf{P}$ as a query vector and compute its compatibility with features in $\mathbf{X}$, resulting in a comparability map $\mathbf{M}$. The feature warping is then computed by the weighted sum of $\mathbf{X}$ over $\mathbf{M}$. PAN has several advantages over previous works. First, PAM operates on every position of the original source feature map in a top-down manner so that the fine details of middle-level features can be kept to reveal the long-range relations between different parts of an object or human. Second, PAM can be applied to multi-granularity to propagate the information from low-level or mid-level features to high-level part features. The compatibility maps show that PAN can focus on distinctive regions while neglecting the background.

The contributions of this paper are two folds. First, a novel module, namely PAM, is proposed, which uses the distinctively learned part features to aggregate information from the source features under the CNN framework. Second, the proposed PAN can propagate useful information from low-level or mid-level features to high-level part features, suppressing background, and distractors while keeping fine-grained details. Our proposed PAN is simple and demonstrates exceptional performance. Our experiments show that PAN significantly boosts the performance of the baseline and achieves leading performance on person ReID.

2 Related Works

Multi-level feature aggregation. It has been found that fusing the feature maps from multiple layers in CNN can result in better feature discrimination ability because features from different layers deliver different levels of seman-

tic information. Generally, features from shallower layers encode image local structures and more fine-grained information, however, they lack global semantic information and often contain noises. Features of deeper layers are believed to contain high-level global semantics but lose spatial and detailed information. Therefore, multi-level feature aggregation has been widely used in many computer vision tasks [24–27, 21].

Multi-level networks can be implemented by using multiple branches on different scales and regions. Li *et al.* [22] proposed a network with multiple branches to fuse local and global features for human feature representation. Qian *et al.* [23] proposed a multi-scale stream layer, which is inspired by GoogLeNet [28] to learn features on different scales. FPN [27] combines feature maps of different scales with element-wise operations. Some methods try to use layers with various scales in a bottom-up manner. For example, SKNet [29] uses multiple branches with different kernel sizes. ACNet [30] adaptively determines connections among feature nodes, which is a general form of connections in CNNs, MLP, or NLN [31].

Attention mechanism for person re-identification. Attention mechanism has been used in many computer vision tasks to regularize the network to focus on essential signals without being interfered with much by outliers. Zhao *et al.* [14] used convolutional layers to learn spatial attention masks for different human parts, which are applied to the feature maps to select the region of interests. In addition to spatial attention, channel attention has also been studied in [22], where the spatial dimension is squeezed and the channel scaling is learned by fully connected layers.

Our work is related to the self-attention [32] and non-local neural networks [31]. Self-attention [32] allows the model to identify multiple locations of the encoded features for machine translation. In [31], a non-local layer is proposed as a basic building block, and a non-local operation is employed to capture spatial and temporal long-range correlations within a feature map. However, the non-local operation has a high time and space complexity when applied to spatially large feature maps. Besides, when the non-local operation is applied to high-level feature maps, the performance gain is marginal because the spatial resolution of high-level feature maps is relatively small. Our work differs from [32, 31] in that we consider the relations between two feature maps on different layers of a CNN.

A top-down attention module is proposed in [33], which takes the final feature representation as a query vector to extract information from the mid-level feature maps. However, the spatial structure of the final feature map is also valuable for fine-grained feature learning. For example, the human body prior is important for person ReID. Simply applying max/avg pooling on the final feature map to aggregate information from mid-level features without considering the spatial structure of the final feature map will limit the model performance. Besides, we use attention modules to learn part features automatically, which utilize the part feature prior to relieve the misalignment issue. Our experiments show that PAN can aggregate the information while retaining the spatial structure of the feature map. Different spatial parts of the final feature map would extract useful information from different parts of the objects.

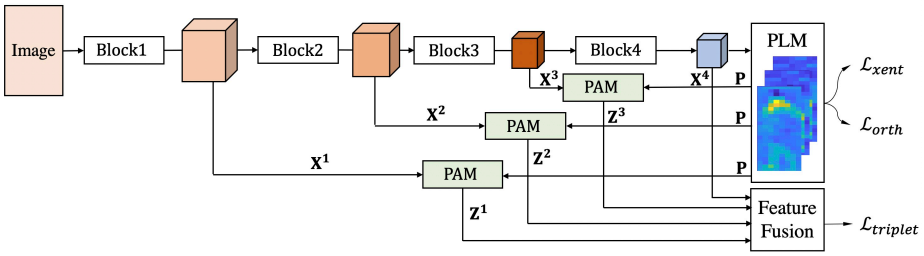


Fig. 1: Illustration of our proposed PAN.

3 Part-aware Attention Network

The motivation of the proposed Part-aware Attention Network (PAN) is to enhance the attentiveness of feature extraction with learned part features. As illustrated in Fig. 1, it contains two modules: part-aware attention module and part feature learning module. The part-aware attention module uses part features as attentiveness guidance and provides strong supervision information for the network at the middle layers. Part feature learning module generates part features by applying both spatial and channel attention to the output of the feature extraction backbone. The output features of PAM are then fused to generate final feature representation. We introduce these modules in the sections below.

3.1 Part-aware attention

The Part-aware Attention Module (PAM) takes part features as guidance for feature learning at the middle layers of the network. We use $\mathbf{P}$ to represent the part feature map, which is the combination of part feature vectors. The details of the part feature learning module would be introduced in the next section. The part-aware attention calculation between a part feature vector $\mathbf{P}_i$ and the middle-level feature map $\mathbf{X}$ can be represented as follows:

$$\mathbf{Z}_i = \frac{\theta(\mathbf{P}_i, \mathbf{X}) f(\mathbf{X})}{N}. \tag{1}$$

where N is the number of total positions in $\mathbf{X}$. PAM takes a target part feature vector $\mathbf{P}_i$ as query and computes its compatibility score with $\mathbf{X}$ using function θ. After computing the compatibility score for each pair of $\mathbf{P}_i$ and $\mathbf{X}$, the feature mapping is computed as a weighted sum over feature map $f(\mathbf{X})$. This process generates a mapping feature $\mathbf{Z}_i$ for every $\mathbf{P}_i$. Function f transforms feature map $\mathbf{X}$ for mapping. The feature map is then normalized with constant N. For implementation of θ, we apply dot-product due to its simplicity and efficiency:

$$\theta(\mathbf{P}_i, \mathbf{X}) = \phi(\mathbf{P}_i)^T \psi(\mathbf{X}), \tag{2}$$

where the function ϕ and ψ transform the features to the same embedding space D. The dimension of D is a hyper parameter, and we would discuss it in the

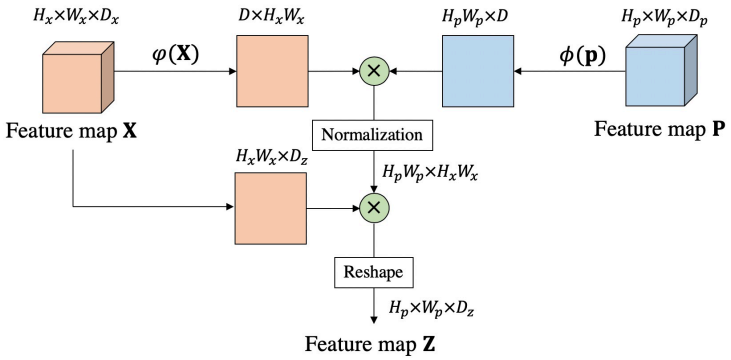


Fig. 2: Our proposed part-aware attention module, which can be applied between part features and middle-level features of a CNN.

ablation study section. It is worth mentioning that when dot-product is applied, the compatibility score computation can be viewed as applying a 1×1 convolution on the feature map $\mathbf{X}$ with kernel $\mathbf{P}_i$.

Combining Eqs. 1-2, we can reformulate the part-aware feature attention module as below:

$$\mathbf{Z} = \phi(\mathbf{P})\psi(\mathbf{X})^T f(\mathbf{X}), \tag{3}$$

where ϕ, ψ and f are convolutional layers, which learn to transform features. The computation process is illustrated in Figure 2. The input feature maps $\mathbf{X}$ and $\mathbf{P}$ are transformed and reshaped so that the computation of compatibility score and feature mapping can be represented as matrix multiplication. The aggregated feature could then be added or concatenated to generate the final feature.

For middle-level feature extraction of PAN, we apply PAM between part features with outputs of the middle blocks of networks. For instance, as shown in Fig. 1, when resnet50 is used as the backbone network, the first three block outputs $\mathbf{X}^l, l \in \{1, 2, 3\}$ and part feature $\mathbf{P}$ are used for PAM. PAN can efficiently take the high-level feature map as the query and gather the feedback information from low-level or mid-level feature maps. The features from mid-level layers of a neural network could provide more part details of the human body in person Re-ID. It is an efficient way for multi-level feature learning, where features from different layers of the network are all mapped to the final feature map. This is more computationally efficient than multi-branch networks for multi-scale feature learning, and is more powerful than vanilla addition or concatenation connection for feature fusion, as it operates on all spatial positions.

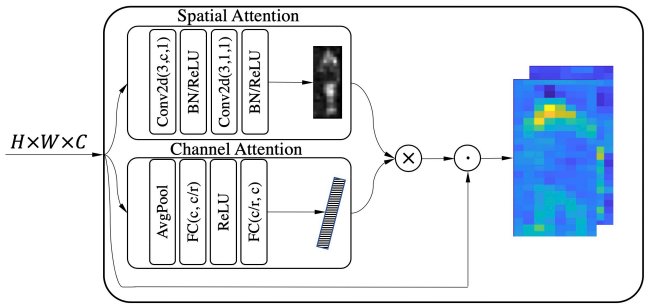


Fig. 3: The proposed part-feature learning module. It contains spatial and channel attention to learn part feature maps.

3.2 Multi-granularity part feature learning

The typical part feature generation process uses priors in humans to model different parts of the human body. For example, human part masks and horizon strips are two popular pooling strategies for learning discriminative human body part representations [5, 14]. Horizon strip [5] is a popular strategy for human part feature generation, due to its simplicity and effectiveness. It divides the final feature map evenly into multiple regions in the height direction. For the feature map of each region, average pooling is applied and the feature is passed into an embedding learning layer. However, this can cause misalignment of the human body due to the differences in the human body in scale, and a person may not be well located at the center of the image. Although, the refined part pooling can be used to reduce the influences of outliers, however, the initial part feature generation can still be affected by the misalignment.

We propose a part-feature learning module (PLM) to learn the part representation automatically. As shown in Fig. 3, the module consists of both channel and spatial attention layers. The spatial attention module consists of a convolutional layer, BN layer, and ReLU activation. The design of the channel attention layer follows [34], where average pooling, fully connected layers, and ReLU activation are used to obtain channel scale vector. To obtain multiple granularities of part features, we propose a multi-granularity part learning strategy. For granularity at scale-i, several i parts are learned automatically by PLM. Here, we use $\mathbf{s}_j^i \in \mathbb{R}^{h \times w}$ to represent the spatial attention mask, where $h \times w$ represents the spatial position in a feature map, (i, j) represents mask for scale i and part $j \in \{1, \ldots, i\}$. Denote by $\mathbf{c}^i \in \mathbb{R}^c$ the channel attention vector. The total attention mask is $\mathbf{m}_j^i = \mathbf{c}^i \otimes \mathbf{s}_j^i \in \mathbb{R}^{c \times h \times w}$. For example, $\mathbf{m^1}$ is the proposed body part at scale 1 ; two distinct part region $\mathbf{m_1^2}, \mathbf{m_2^2}$ are learned at scale 2 to divide the feature map into two parts. As shown in Fig. 1, given the $\mathbf{X^4} \in \mathbb{R}^{c \times h \times w}$, which is generated by the backbone feature extractor, the part feature vector can be computed as $\text{AvgPool}(\mathbf{m}_j^i \odot \mathbf{X^4})$. The final part feature map $\mathbf{P}$ is the combination of all part feature vectors. All the part features are trained with cross-entropy loss, and we use orthogonal regularization to learn distinct part-feature maps.

This multi-granularity part feature learning strategy can learn different granular part features and boost learning efficiency.

3.3 Loss functions

In addition to the classical cross-entropy loss [35], denoted by L_{xent}, we also employ the triplet loss and an orthogonal regularization loss to train our models. For triplet loss, we adopt the online hard triplet mining strategy proposed in [36], which considers the hardest triplets within a mini-batch.

To enforce the learned part regions to be distinctive, we add an orthogonal regular term to reduce the overlap of different part masks:

$$\mathcal{L}_{orth} = \frac{1}{L} \sum_{l \in \{1,...L\}} \|\mathbf{M^l}\mathbf{M^l}^T - \mathbf{I}\|_{\mathbf{F}}, \tag{4}$$

where $\mathbf{M^l}$ is the part region masks at scale-l and each row of it represents a part region. L is the total number of scales. $\mathbf{I}$ is an identity matrix.

We apply cross-entropy loss L_{xent} on every part features generated by part feature learning module. The triplet loss $L_{triplet}$ is applied to the final feature representation (i.e., the concatenation of all part features). The final loss function is:

$$\mathcal{L} = \mathcal{L}_{xent} + \lambda_1 \mathcal{L}_{triplet} + \lambda_2 \mathcal{L}_{orth}, \tag{5}$$

where $\mathcal{L}_{xent}$ is the sum of all the part cross-entropy loss functions, $\mathcal{L}_{triplet}$ is the triplet loss for the final feature and λ_1 and λ_2 are trade-off parameters.

3.4 Discussion

Multi-level feature fusion. Most current multi-level feature fusion methods use element-wise operations such as addition or concatenation to fuse feature maps from different layers. Since the channel and spatial dimensions of the feature maps are different, usually a function f is first applied to downsample and reshape the source feature map, and then a function g is applied to fuse the source and target feature maps. Both f and g are usually implemented by convolutional layers. Though simple to implement, the downsides of this process are two folds. First, the downsampling function f would reduce the spatial dimensions of the feature map and lose fine-scale information contained in the feature maps. Second, the fusing function g neglects the long-range relations between two feature maps, as convolution only operates within a local area. The proposed PAN overcomes these problems by considering fine-grained pair-wise relations between feature maps.

Comparison with other attention based methods. Some previous works on Re-ID have also utilized attention modules in feature extraction. For example, HAN [22] utilizes spatial and channel attention in the middle layers of the network. Both SCAN [37] and IANet [38] use self-attention module in the network design, which is similar to non-local (NL) [31] module. Spatial attention, channel

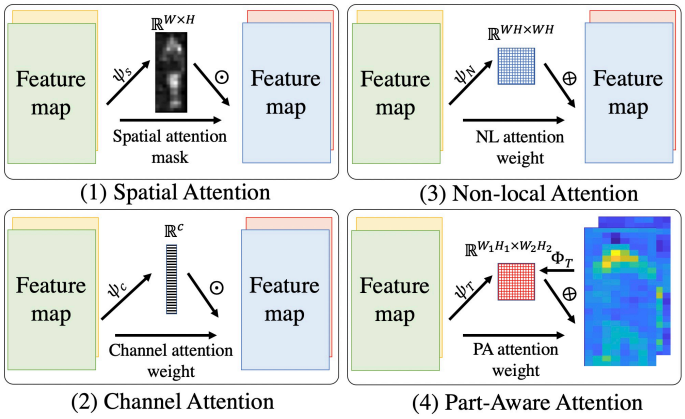


Fig. 4: Comparison of part-aware attention with other attention methods. (1) *Spatial attention* applies convolution and softmax functions to produce a spatial mask. (2) *Channel attention* uses fully connected layer to generate a scale vector. (3) *Non-local attention* uses self-similarity to generate a self attention matrix. (4) Our *part-aware attention* generates part-aware attention map for different human parts.

attention, and self-attention can all be viewed as bottom-up attention methods, while our proposed PAM is a top-down attention module. Fig 4 compares PAM with other attention methods. We can see that PAM uses both source feature map and target feature map to compute the attention maps of different parts, while other methods only use source feature map for attention computation.

PAM is related to NL modules but they have significant differences. NL modules capture long-range relations within a feature map, while PAM operates across different feature maps. To map the low-level or mid-level features to high-level part features, PAM takes the part features learned from deeper layers (often have more high-level semantic information) as a query to explore the discriminative information from shallower layers of a network. Besides, due to the reduced spatial resolution along with the feed-forward process in CNN, the computational cost of PAM is significantly less than the vanilla self-attention methods or NL methods.

4 Experiments on person ReID

4.1 Implementation details

Data preprocessing. We use the same data pre-processing methods on all datasets. In the training stage, common data augmentation methods are applied to images, including random flipping, shifting, zooming, cropping, random erasing [39]. The images are then resized to 384×128.

Backbone network. We use ResNet50 as the backbone network since it has been successfully used in person Re-ID. Specifically, we adopt the modified ResNet50 architecture in [5], in which the stride of last convolutional layer is set to 1 to benefit final feature learning. We use a 1×1 convolutional layer for embedding learning and set output dimension to 256 for each embedding layer.

Settings of PLM and PAM. The PLM contains both spatial and channel attention. The spatial attention contains two stacked Conv2D-BN-ReLU blocks, as shown in Fig. 3. The design of channel attention follows SE-Net [34], which consists of an average pooling layer, a fully connected layer and a ReLU. The two attention maps are then combined to generate the part features. The PAM aim to transform input feature maps to the same embedding space. We first use a 1×1 convolutional layer to transform the output dimension of the feature maps according to the fusion method. For addition and concatenation based feature fusion, we set the output dimension of feature map $\mathbf{Z}$ in Eq. 3 equals to the dimension of $\mathbf{P}$ and $\mathbf{X}$, respectively.

Optimization. Our model is trained by randomly selecting 8 identities with 4 samples each identity as a batch. The SGD optimizer is employed. The learning rate is set to 1×10^{-1} for the parameters of PAN, embedding layer and softmax classification layer, and the rate is set to 1×10^{-2} for the pre-trained parameters of the network. The learning rate is divided by 10 after 5000 and 7500 iterations, and the training is stopped after 10,000 iterations.

4.2 Ablation study

The number of granularity in PLM. In this section, we study how the number of granularity affects the performance of our multi-granularity PLM method and compare it with the horizon strip pooling method [5]. Each pooled part feature is followed by an embedding learning layer and a softmax cross-entropy loss function. We test 3, 6, and 8 parts of horizon stripe, as well as four different granularity. All the features are then passed to an individual embedding layer and softmax loss function. The final person feature representation is the concatenation of all part features. For the simplicity of experiments, we set $\lambda_1 = 1$, $\lambda_2 = 10^{-3}$ throughout the experiments.

The experimental results are listed in Table 1. Part feature learning module (PLM) performs better than horizon strip pooling when they have the same total part number. For instance, three granularity scales PLM (combined with scale 1, 2 and 3) and 6-part horizon pooling have the same part number and final feature dimension, while three scales PLM outperforms 6-part horizon strip pooling by 0.8% on mAP. Using more than three scales will not further improve the performance of PAM. As we use orthogonal regularization term to force the learned attention maps to be distinct, using too many granularity scales may divide the human body into too many parts, and deteriorate the performance. We empirically found that using three granularity scales leads to the best results. For simplicity, we use PLM with three scales by default in the following experiments.

Table 1: Results of different part feature extraction strategies on Market1501.

Method		Top1	Top5	Top10	mAP
Global Pooling		86.0	94.3	96.2	68.9
Horizonstrips	3 parts	91.7	96.8	97.9	78.1
	6 parts	92.5	96.8	98.1	79.5
	8 parts	92.4	97.0	98.0	79.2
Multi-granularity	$\{1\}$	90.0	96.4	97.4	75.4
	$\{1,2\}$	91.8	96.4	97.5	78.9
	$\{1,2,3\}$	**92.9**	**97.4**	**98.4**	**80.4**
	$\{1,2,3,4\}$	92.8	96.9	98.1	79.9

Table 2: The effectiveness of PAN on Market1501. *Baseline* stands for the model trained without the multi-layer connection.

Model	D	Top1	mAP	Param	Memory	Time
Baseline	-	93.0	81.1	27.8M	17.79G	72s
Baseline_w	-	93.4	81.8	29.2M	17.80G	74s
Conv_a	-	93.5	82.0	35.2M	17.83G	76s
Conv_c	-	93.7	82.5	32.3M	17.81G	75s
PAN_a	128,128,128	93.5	83.0	32.5M	17.81G	75s
	256,256,256	93.8	83.5	33.5M	17.82G	77s
	256,512,1024	93.9	83.6	36.6M	17.84G	79s
PAN_c	128,128,128	93.8	83.7	33.0M	17.82G	76s
	256,256,256	94.3	84.2	33.9M	17.83G	76s
	256,512,1024	94.3	84.1	37.0M	17.85G	78s
NL	128,256,512	93.9	83.9	30.5M	21.63G	79s
PAN_c + NL	128,256,512/256	94.8	85.3	36.7M	21.66G	84s

Feature fusion in PAN. We investigate how to fuse the transformed features in the proposed PAN. The baseline is a plain ResNet50 network with PLM and trained with triplet loss and softmax loss. PAN_a and PAN_c stand for applying PAN and fusing features with addition or concatenation. We use *Conv* connection blocks Conv_a and Conv_c for better comparison, which contains convolution ($kernel = 1, stride = 2$), batch nomalization and ReLU. For convolution, the channel dimension is gradually increased to 2048 for addition connection ($256 \rightarrow 512, 512 \rightarrow 1024, 1024 \rightarrow 2048$), and retain the same for concatenation.

In Table 2, the models trained with feature mapping (PAN_a and PAN_c) improve the performance over baseline network by 1.9% and 2.6% on mAP when the inter-channel number D is 128. Increasing the number of inter-channels D can further increase the performance. We choose inter-channel to be 256 for all the three blocks by considering the trade-off between parameter number and performance. The performance gain by both connections (PAN_a and PAN_c) over Conv_a and Conv_c provides evidence that PAN leverage extra information, and

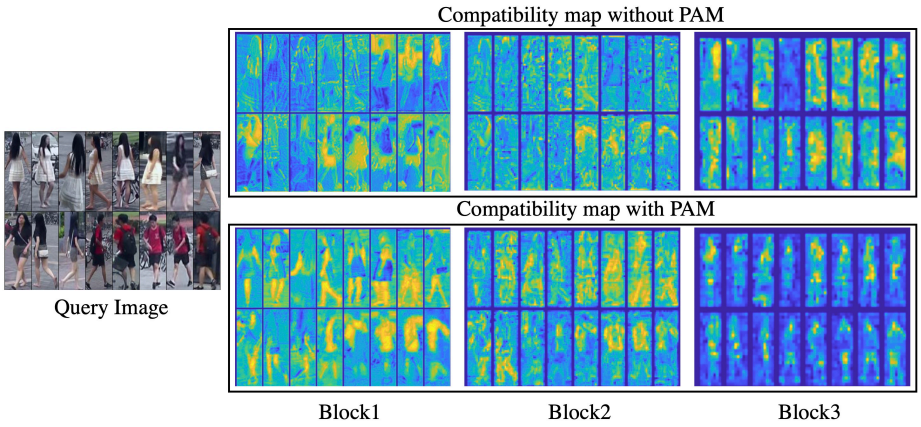


Fig. 5: Compatibility maps of PAM for different blocks of ResNet50. The compatibility map is computed by using the first scale part feature vector of PLM against block1-3 (ref to Eq. 2). The strength of yellow color indicates the degree of compatibility. The top right panel shows the compatibility maps without PAM, which are vague and unclear. The bottom right panel shows the compatibility maps with PAM. PAM learns to focus on discriminative parts and can capture long-distance relations.

it is not simply due to the increase of number of parameters, as PAN_a and PAN_c have less parameters than Conv_a and Conv_c and faster at inference. We use feature maps of the third stage as the local feature and last stage as global feature to construct the Baseline_w, it performs better than using global feature alone but lower than our proposed method. The test time is measured by using 1 Quadro GV100 GPU. The results show that PAN has small extra computational cost and it is efficient in practice.

Benefit of orthogonal regularization term. We conducted experiments to investigate the role of orthogonal regularization. The parameter is $\{0, 10^1, 10^2, 10^3, 10^4\}$. The mAP are 80.36%, 80.21%, 80.87%, 81.14%, 80.46%, respectively. We thus choose 103 in our experiment One can see that the benet over no orthogonal regularization is 0.78%.

Visualization of PAM. Figure 5 shows the compatibility map (computed using Eq. 2) of the first scale part feature vector from PLM against feature maps of first three blocks. The yellow color indicates the strength of the compatibility. To better illustrate the advantages brought by proposed method, in the top right panel of Figure 5 we show the compatibility maps without using PAM, which are vague and unclear. The bottom right panel shows the compatibility maps with PAM. We can see that PAM learns to capture the correlations between outputs of different blocks, and propagate information from mid-level and low-level feature maps to high-level features to enhance the discriminative parts of human body. The compatibility maps of first two stages focus on color and texture. While

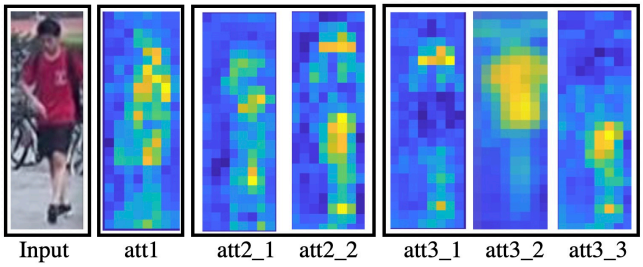


Fig. 6: A closer look of the compatibility maps of PAM.

for the third stage, the compatibility maps represent different human parts and their combinations. This observation indicates PAN learns distinct semantics mapping from different stages of networks.

In Figure 6, we provide a closer look of PAM at different granularity of the third stage. For the first granularity scale, the attention map focus on the whole body. In the second granularity scale, the body is roughly divided into two parts. While in the third granularity scale, the network learns to divide the body into three parts. As we can see, all the parts are quite different at each granularity scale, which indicates PLM's ability to learn distinct body part feature.

Non-local module and PAN. As we mentioned in the related work section, the non-local (NL) module used in the previous work [31] is different from our proposed PAN. NL module captures long-range relations within a feature map, while our proposed PAN operates across different feature maps. NL module can be complementary to our PAN module according to their different functionalities.

We conduct experiments to add NL modules after block1-3. Dot product is used the compatibility computation and the inter-channel is set to half of the original channels. We also downsample the feature map by default. The training memory consumption is tested with batch size 128. The result is in Table 2. Since the output resolution of block4 is much smaller than earlier blocks (4 times to 16 times smaller in ResNet50), PAN is much faster and costs less memory than NL for compatibility computation. NL module can improve the performance over *baseline* but with a higher computation and memory cost. Combining NL with PAN can further improve the performance by 0.9%/1.4% on Top1/mAP.

4.3 Comparison with state-of-the-arts

In this section we compare our PAN_c, denoted by PAN for simplicity in the following parts, with state-of-the-art methods on the three benchmark person Re-ID datasets. We adopt BNNeck in [47] to further boost the performance of our model. All the reported results are single-query without re-ranking.

Market1501. Market1501 is one of the largest benchmark datasets for Person Re-ID, and many methods have been reported on this dataset. We compare the proposed method with most of the state-of-the-arts. The experimental results are shown in Table 3. With ResNet50 as the pre-trained network, the

Table 3: Single-shot performance comparison of different methods on Market1501. Methods with * are attention methods and with # are fusion methods.

Method	Top1	Top5	Top10	mAP
Spindle# [16] (CVPR17)	76.9	91.5	94.6	-
DLPA [14] (ICCV17)	81.0	92.0	94.7	63.4
SSM [40] (CVPR17)	82.2	-	-	68.8
TriNet [36] (Arxiv17)	84.9	94.2	-	69.1
MLFN# [3] (CVPR18)	90.0	-	-	74.3
HA-CNN* [22] (CVPR18)	91.2	-	-	75.7
DuATM* [4] (CVPR18)	91.4	97.1	-	76.6
PCB [5] (ECCV18)	93.8	97.5	98.5	81.6
MGN# [19] (MM18)	95.7	-	-	86.9
Local CNN* [41] (MM18)	95.9	-	-	87.4
IANet* [38] (CVPR19)	94.4	-	-	83.1
Pyramid# [42] (CVPR19)	95.7	98.4	99.0	88.2
MHN* [43] (ICCV19)	95.1	98.1	98.9	85.0
ABD-Net* [44] (ICCV19)	95.6	-	-	88.3
SONA* [45] (ICCV19)	95.6	98.5	99.2	88.8
ST-ReID [46] (AAAI19)	97.2	99.3	99.5	86.7
PAN*	96.0	98.6	99.3	89.0

proposed PAN approach achieves 89% mAP and 96.0% CMC top1. PAN outperforms PCB [5] by 2.2% on CMC top1 and 7.4% on mAP, which applies horizon part pooling and refine part pooling on the final feature map, on Market1501. Comparing to multi-branch methods such as Spindle [16], MLFN [3], HA-CNN [22] and MGN [19], PAN is a single branch ResNet50 with target aware mid-level feature connections and much less parameters. It surpasses the MGN [19] on all three datasets. PAN outperforms Local CNN [41], which fuses local and global features in the mid-level of CNN with Local CNN module. It also outperforms state-of-the-art local-global fusion method Pyramid [42] and self-attention method IANet [38]. Comparing with recent bottom-up attention methods SONA [45], ABD-Net [44] and MHN [43], our method considers both bottom-up and top-down attention and achieves slightly better performance. This indicates the effectiveness of PAN, which use the part feature as guidance to fully utilize the mid-level features of network.

CUHK03. On CUHK03, we follow the new protocol proposed in [48] and conduct experiments using labeled datasets. The results are shown in Table 4. PAN achieves 82.5% CMC top1 and 80.4% mAP on CUHK03 labeled dataset, respectively, which are leading results on the CUHK03 dataset. PAN exceeds DaRE [49], which applies deep supervision to the mid-level features without considering the spatial structure of feature maps. It also outperforms multi-branch attention method CAMA [50] with much less parameters.

DukeMTMC-ReID. On this dataset, we compare our method with all the state-of-the-art methods in the literature. As shown in Table 5, our model

achieves much better performance than other methods. The proposed PAN obtains 89.5% Top1 accuracy and 79.2% mAP, respectively. PAN beats several attention and multi-scale based methods, including HA-CNN [22] (harmonious attention and local features of every building block are extracted and processed), IANet [38] (self-attention embedded in the middle of the networks), Pyramid [42] (multi-loss and pyramidal model to incorporate local and global information), SONA [45] (Second-order attention with dropblock), ABD-Net [44] (Channel and Position attention) and MHN [43] (mixed high-order attention). PAN efficiently suppresses the background noise and utilize the useful middle-level features. The strong performance shows that PAN is a promising direction to utilize attention for feature fusion.

Table 4: Single-shot performance comparison on CUHK03 dataset.

Method	Top1	mAP
SVDNet [1] (ICCV17)	41.5	37.3
HA-CNN* [22] (CVPR18)	44.4	41.0
MLFN# [3] (CVPR18)	54.7	49.2
DaRE [49] (CVPR18)	58.1	53.7
Local CNN* [41] (MM18)	58.7	53.8
MGN* [19] (MM18)	68.0	67.4
CAMA [50] (CVPR19)	70.1	66.5
MHN* [43] (ICCV19)	77.2	72.4
SONA* [45] (ICCV19)	81.4	79.2
PAN*	82.5	80.4

Table 5: Single-shot performance comparison on DukeMTMC-ReID.

Method	Top1	mAP
HA-CNN* [22] (CVPR18)	80.5	63.8
Local CNN* [41] (MM18)	82.2	66.0
PCB [5] (ECCV18)	83.3	69.2
MGN# [19] (MM18)	88.7	78.4
IANet* [38] (CVPR19)	87.1	73.4
MHN* [43] (ICCV19)	89.1	77.2
SONA* [45] (ICCV19)	89.4	78.3
ABD-Net* [44] (ICCV19)	89.0	78.6
ST-ReID [46] (AAAI19)	94.0	82.8
PAN*	89.5	79.2

5 Conclusions

We proposed a simple yet effective part-aware attention network to leverage and strength multi-layer features in convolutional neural networks. The so-called part-aware attention network (PAN) connects source features to the target features and learns to leverage their pair-wise correspondence for feature enhancement. It considers not only the local spatial relations of multi-layer feature maps but also the long-range relations among them. In person Re-ID, PAN provides an effective way to facilitate the interaction between low-level/middle-level features and high-level features to strength the discrimination of human body. Detailed analysis and extensive experiments were conducted on three widely used datasets to validate the effectiveness of our PAN approach for person Re-ID.

Acknowledgements:

This research is supported by the China NSFC grant (no. 61672446).

References

1. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: ICCV. (2017)
2. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: CVPR. (2017)
3. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: CVPR. Volume 1. (2018) 2
4. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: CVPR. (2018)
5. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV. (2018)
6. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Part-aligned bilinear representations for person re-identification. In: ECCV. (2018)
7. Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., Li, S.Z.: Embedding deep metric for person re-identication A study against large variations. In: ECCV. (2016)
8. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR. (2015)
9. Jose, C., Fleuret, F.: Scalable metric learning via weighted approximate rank component analysis. In: ECCV. (2016)
10. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR. (2016)
11. Liao, S., Li, S.Z.: Efficient psd constrained asymmetric metric learning for person re-identification. In: ICCV. (2015)
12. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR. (2017)
13. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR. (2016)
14. Zhao, L., Li, X., Wang, J., Zhuang, Y.: Deeply-learned part-aligned representations for person re-identification. In: ICCV. (2017)
15. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR. (2017)
16. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR. (2017)
17. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. In: CVPR. (2017)
18. Zhang, Y., Li, X., Zhao, L., Zhang, Z.: Semantics-aware deep correspondence structure learning for robust person re-identification. In: IJCAI. (2016)
19. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning Discriminative Features with Multiple Granularities for Person Re-Identification. ArXiv e-prints (2018)
20. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision. (2017) 1–9
21. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: The IEEE International Conference on Computer Vision (ICCV) Workshops. (2017)

22. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR. Volume 1. (2018) 2
23. Qian, X., Fu, Y., Jiang, Y., Xiang, T., Xue, X.: Multi-scale deep learning architectures for person re-identification. CoRR **abs/1709.05165** (2017)
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
25. Yair, N., Michaeli, T.: Multi-scale weighted nuclear norm image restoration. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018)
26. Branson, S., Beijbom, O., Belongie, S.: Efficient large-scale structured learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 1806–1813
27. Kirillov, A., Girshick, R., He, K., Dollar, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
28. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. (2015)
29. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. (2019)
30. Ding, X., Guo, Y., Ding, G., Han, J.: Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: The IEEE International Conference on Computer Vision (ICCV). (2019)
31. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. CVPR (2018)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems 30. Curran Associates, Inc. (2017) 5998–6008
33. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.S.: Learn to pay attention. ICLR (2018)
34. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. (2018)
35. Bishop, C.M.: Pattern recognition and machine learning. (2006)
36. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
37. Zhang, R., Sun, H., Li, J., Ge, Y., Lin, L., Luo, P., Wang, X.: SCAN: self-and-collaborative attention network for video person re-identification. CoRR **abs/1807.05688** (2018)
38. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9317–9326
39. Zhun Zhong, Liang Zheng, G.K.S.L.Y.Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
40. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. (2017)
41. Yang, J., Shen, X., Tian, X., Li, H., Huang, J., Hua, X.S.: Local convolutional neural networks for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference, ACM (2018) 1074–1082
42. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 8514–8522
43. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 371–381

44. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 8351–8361
45. Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 3760–3769
46. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. (2019) 8933–8940
47. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Transactions on Multimedia (2019) 1–1
48. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. (2017)
49. Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., Huang, G., Hariharan, B., Weinberger, K.Q.: Resource aware person re-identification across multiple resolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8042–8051
50. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: 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. (2019) 1389–1398