Pyramid and Similarity Based Feature Enhancement Network for Person Re-identification

Chengguo Chu¹,a, Meibin Qi¹,*b, Jianguo Jiang¹,c, Cuiqun Chen¹,d and Jingjing Wu¹,e

¹DSP Laboratory, School of Computer and Information, Hefei University of Technology, Hefei, Anhui, 230009, China
achengguochu@mail.hfut.edu.cn; *qimeibin@hfut.edu.cn; cjiang@hfut.edu.cn; dchencuiqun_hfut@163.com; hfutwujingjing@mail.hfut.edu.cn;

Abstract. Person re-identification (Re-ID) has become an increasingly important task due to its wide applications in the field of intelligent video surveillance. However, because of the camera shooting angle, the different proportions and pose changes of pedestrians in images are also different, such problem substantially hinders the model’s capability on further improving feature extraction. In this paper, we propose a new end-to-end architecture named Pyramid and Similarity Based Feature Enhancement Network (PS-Net) which includes Pyramid Joint Attention(PJA) and Similarity Feature Fusion Branch(SFF) to enhance the feature maps performance. Among them, PJA contains spatial attention at different receptive fields and channel attention to enhances the discriminative area of pedestrian features by combining the advantages of two different attentions. SFF fuses the features which have different levels of information but have congruent relationships, enriches the information contained in pedestrian features and generates more robust feature. Extensive experiments have been conducted to validate the superiority of our PS-Net for Re-ID over a wide variety of state-of-the-art methods on two large-scale datasets: Market-1501 and DukeMTMC-ReID.

1. Introduction
Person re-identification (Re-ID) aims to correlate the identity of pedestrians at different time and places under different cameras. Given a query image and a large set of gallery, Re-ID uses a feature embedding to represent each pedestrian identity, and sorts each image in the gallery based on the similarity of feature embedding. In recent years, with the Convolutional Neural Network (CNN) has been widely used in more and more fields, more powerful representations have been introduced [1,2] which have significantly improved the performance of Re-ID. However, in practical application scenarios, Re-ID is still a challenging task due to the occlusion, background noise, different proportions, pose changes, perspective changes, and spatial misalignment, etc.

Affected by the aforementioned factors, the feature representations of pedestrian images is actually not good enough. In order to enhance the feature representations, many recent works [3,4,5,6,7,8] use attention mechanism to highlight the informative parts of convolution responses and suppress the noisy patterns. Chen et al.[6] introduced high-order statistics information into attention mechanism, learning the subtle differences among pedestrians and generates the discriminative attention proposals. Chen et al.[7] proposed a novel attention modules which can seamlessly integrates into network and a diversity regularizations method to learn features that are more robust. Li et al.[8] designed a multi-
task learning model that combine the soft pixel-level and hard region-level attention mechanism to generate more discriminative feature representations. However, most methods based on the attention mechanism only focus on feature enhancement at a single receptive fields which is difficult to capture more context information when pedestrians have different proportions in images. As shown in figure 1, the image of (1) and (3) has a large percentage pedestrians parts that the receptive field of a 3×3 convolution kernels maybe obtain information on top or bottom parts but the same size receptive field can get more context information even the whole person body in (2) and (4). Furthermore, these works do not consider the complementary relationship between spatial attention and channel attention.

![Different pedestrian images](image)

**Figure 1.** Different pedestrian images were selected from Market-1501, in which (1) and (2), (3) and (4) were the same pedestrian identity, but the proportion of body parts in the image was different.

In recent year, feature fusion has been proved to be effective in various fields. Zhao et al. [9] designed a human body region guided multi-stage feature decomposition and feature fused by tree-structured competitive architecture, taking human body structure information into feature learning. Lin et al.[10] proposed a top-down method to combine multi-scale features. Ghiasi et al. [11] designed a neural architecture search to automatically design feature network topology. These methods do not take into account that adding extra information will inevitably lead to errors in estimation. Moreover, the low-level features contain less information but more noise, fuse these feature not only increases the amount of computation and complexity of the network, but also difficult to achieve the expect improvement in network performance.

Therefore, a new end-to-end architecture named Pyramid and Similarity Based Feature Enhancement Network(PS-Net) has been proposed, which contains Pyramid Joint Attention(PJA) and Similarity Feature Fusion Branch(SFF) to enhance the baseline learned feature maps. Among them, PJA combines Pyramid Spatial Attention(PSA) and Channel Attention(CA) to take full advantage of the two different attention mechanisms, learn more context information, alleviates the mismatch of information obtained by pedestrians occupying different proportions in images. SFF fuses the similarity feature which have the same spatial size so that the features have a one-to-one correspondence relationships to enrich the features information and enhance the Re-ID performance.

2. The Proposed Method

2.1. Framework

In this paper, we propose a new end-to-end network named Pyramid and Similarity Based Feature Enhancement Network(PS-Net) which contains Pyramid Joint Attention(PJA) and Similarity Feature Fusion Branch(SFF). Note that the stride of the spatial down-sampling operation of the last layer is set to 1, so that the conv4 output feature and conv5 output feature of the baseline have same spatial size. The overview framework as shown in figure 2.

In the whole network structure, we embed PJA behind conv2, conv4 and conv5 of the baseline to strengthen the outputs features. After conv5, the network is divided into Global Branch(GB) and Similarity Feature Fusion Branch(SFF). GB extracts the conv4 and conv5 features after PJA, guide the PJA to reinforce the discriminative region and improve baseline performance. SFF fuses the conv4 and conv5 feature after PJA, enrich the information contained in pedestrian feature. Note that we use Global Maximum Pooling(GMP) and Global Average Pooling(GAP) as our pooling strategy. During
the training phase, we use the corresponding supervised signal for the feature of conv4 and conv5 after PJA and the features of the SFF output, respectively.

Figure 2. The framework of PS-Net. 1×1 refers to 1×1 convolution. The network is split into Global Branch and Similarity Feature Fusion Branch after conv5. The conv4 and conv5 output features are concatenate as a fused feature sent to Similarity Feature Fusion Branch. Best viewed in color.

2.2. Pyramid Joint Attention

In order to combine the advantages of two different attentions. We propose a new Pyramid Joint Attention(PJA) that include Pyramid Spatial Attention(PSA) and Channel Attention(CA), the architecture is shown in figure 3. Input feature is a 3-d tensor \( F \in \mathbb{R}^{H \times W \times C} \).

Figure 3. Illustration of Pyramid Joint Attention. The input feature is sent to two branches respectively for global spatial attention and channel attention. Notice that the final attention map will multiply with input feature element by element as output feature.

The PJA module first utilize two 1×1 convolution on the input feature to reduce the channel dimension, respectively, and get two output features \( I_1 \in \mathbb{R}^{H/2 \times W/2 \times C} \), \( I_2 \in \mathbb{R}^{H/2 \times W/2 \times C} \). Then, \( I_1 \) and \( I_2 \) are sent to PSA and CA respectively for global attention operation to obtain the attention maps \( A_1 \) and \( A_2 \). In order to integrate the attention output feature, we concatenate the output attention maps \( A_1 \) and \( A_2 \) to form a fused attention map \( A \in \mathbb{R}^{H \times W \times C} \), a 1 × 1 convolution has been used to integrate information of A. Inspire by [5], we add 1.0 to Sigmoid function, the function can formulate as:

\[
Y = \frac{1}{1+e^{-Z}}
\]

(1)

Where \( Z \in \mathbb{R}^{H \times W \times C} \) is the output for 1 × 1 convolution, \( Y \) is attention score map. Finally, we multiply the attention score map \( Y \) and the input features element by element to obtain the output feature map \( F \in \mathbb{R}^{H \times W \times C} \). The formula of F is as follows:

\[
F = Y \times I
\]

(2)
Where $\times$ denote element-wise multiplication.

**Figure 4.** Illustration of Pyramid Spatial Attention. The input feature will adopt four different kernels size $1 \times 1$, $3 \times 3$, $5 \times 5$, $7 \times 7$ respectively, and concatenate four attention maps as the final spatial attention map. Best viewed in color.

2.2.1 Pyramid Spatial Attention. In consideration of the convolutional receptive field on a single scale cannot well contain more context information and lead network to miss some detailed information. A new Pyramid Spatial Attention (PSA) has been proposed to learn the pedestrian features under different receptive fields. As shown in figure 4, PSA uses the convolution operations of four different convolution kernels to capture more global context information under different receptive fields. In our experiment, we set the size of convolutions kernels to $1 \times 1$, $3 \times 3$, $5 \times 5$, $7 \times 7$ with the padding size of 0, 1, 2, 3, respectively. Since the resolution of high-level feature maps is small, using large kernel size does not bring too much computation burden. The spatial attention output $A_1 \in \mathbb{R}^{H \times W \times C/2}$.

2.2.2 Channel Attention. In [12], a novel Squeeze-and-Excitation (SE) block has been proposed, extensive experiments has been proved that the SE block can effectively improve the feature performance and convenient to plug into various network structures. Inspired by this, we adopts this method as our Channel Attention (CA). As shown in figure 5. Firstly, the Global Average Pooling (GAP) and Global Maximum Pooling (GMP) on each channel of the input feature, respectively, and the two output were added as the pooled results, as shown below:

$$P = \frac{1}{H \times W} \Sigma^H_{j=1} \Sigma^W_{i=1} c_i \left( i, j, \frac{1}{2} \right) + \max \left( c_i \left( :, i, \frac{1}{2} \right) \right)$$

(3)

Where $P$ is the output of pool operation. Then, we send the pooled results $P$ to a SE block. In order not to destroy the spatial structure of the feature map, $1 \times 1$ convolution operation is used in our experiment instead of the fully connected layer. The output of the channel attention is $A_2 \in \mathbb{R}^{H \times W \times C/2}$.

**Figure 5.** Illustration of Channel Attention. The input feature first adopted with GMP+GAP pool and then the squeeze-excitation is used to generate final channel attention map.
2.3. Global Branch
This branch is used to train the conv4 and conv5 output features after PJA, guide the PJA to focus on the informative parts and generate more discriminative features. In this branch, the GAP + GMP pooling operation is used on the output features of the conv4 and conv5 respectively. Then, the dimensions of the two features is reduced to 512-d by a $1 \times 1$ convolution. Inspired by [13], we add the BNNeck technique before classification to balance the various dimensions of the classification vector so that triplet loss and cross entropy loss can converge simultaneously.

2.4. Similarity Feature Fusion Branch
In this Branch, we first concatenate the conv4 output feature $F_1$ and conv5 output feature $F_2$ to form a fused feature $F_3$. In order to integrate the fused feature information, a $1 \times 1$ convolution is used and reduced the feature to 2048-d. Then, a CNN module is used as feature extract model. In our experiment, the CNN is a simple SE[12] module that contains two $1 \times 1$ convolutions and a $3 \times 3$ convolution. The features are also pooled by GAP + GMP operation. Finally, we use a $1 \times 1$ convolution reduce the channel dimensions to 512-d for center loss[13] and triplet loss[14], the final classification prediction vector is obtained by BNNeck and a classification fully connected layer.

2.5. Loss function
In order to get better training results, during the training phase, we use label smoothing cross entropy[15] for classification loss, and the triplet loss with center loss is used to improve the effect of training. In our experiment, the total loss can be defined as:

$$ L = \sum_{i=1}^{3} \sum_{j=1}^{N} \frac{1}{N} \log(p_j) + \max(d_p - d_n + \alpha, 0) + \frac{1}{2} \sum_{k=1}^{N} \left\| f_{t_k} - c_{y_k} \right\|^2$$

(4)

Where $p_j$ as ID prediction logits of class j, $q_j$ is the probability score of the class j after label smoo-thing, N is the number of training samples, $d_p$ and $d_n$ indicated the positive sample pair distance and the negative sample pair distance, respectively, $\alpha$ is the boundary value in the triplet loss, the $c_{y_k}$ is the $y_k$-th class center of deep features and the $f_{t_k}$ denotes the feature output of the network.

3. Experiment

3.1. Datasets
Datasets: Two widely person re-ID datasets Market-1501[16] and DukeMTMC-ReID [17] are adopted as experimental evaluations. Market-1501 includes 32,668 pedestrian images with 1501 identities which captured from 6 different cameras. For DukeMTMC-reID, it consists of 36411 images of 1812 identities, These pedestrian images are catched by 8 camera views which are not in the overlapped region of a university. In order to evaluate PS-Net performance, the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) has been used in our experiment.

3.2. Implementation
During the training phase, the input image is resized to $384 \times 128$ and enhanced by random horizontal flipping, normalization and random erase. The test image was resized to $384 \times 128$ and only normalized. Baseline uses IDE[18] network which feature extraction module is the pretrained Se-Resnext50[12] on the ImageNet dataset. The boundary value $\alpha$ of the triplet loss is set to 1.2 and the weight $\beta$ of the center loss is set to 0.0005. The SGD with the momentum is 0.9 and weight decay set to 0.0005 has been used as our optimizer. Our proposed model trains a total of 200 epochs, and the initial learning rate is set to 0.02, which is reduced to 0.002 and 0.0002 at 80 epoch and 140 epoch, respectively. In the test phase, we combine all 512-d features $F_1$, $F_2$ and $F_3$ into 1536-d feature to predict. All our experiments on different datasets follow the settings above.
3.3. **Performance Comparison**

**Market-1501:** In this dataset, the comparison results with state-of-the-art methods are shown in table 1. The results show that PS-Net has reached better performance with Rank1/mAP=95.9%/89.7%. Compared to ABD-Net[7], our method has improved 1.6% on mAP and 0.3% on Rank-1. RGA [3] is the best current results on this dataset, compare with this method, our method improves 1.6% on mAP and 0.1% on Rank-1. Our method also improves 1.5% on mAP and 0.2% on Rank-1 compare with Pyramid[19].

**DukeMTMC-reID:** According to table 1, our PS-Net achieves the effect of Rank1/mAP =90.4%/80.3% and surpasses the MHN[6] which uses a high-order mixed attention mechanism by 3.1% and 1.2% on mAP and Rank-1. Compared with SCSN[20], it still improves 1.2% and 0.2% on mAP and Rank-1 respectively. It also has 2.8% mAP and 0.2% Rank-1 improvement compared with MGN[21].

![Table 1. Performance(%) on Market-1501 and DukeMTMC-reID dataset.](image)

| Methods    | Market-1501 | DukeMTMC-reID |
|------------|-------------|---------------|
|            | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP |
| Ours       | 95.9   | 98.4   | 89.7 | 90.4   | 95.2   | 80.3 |
| RGA-SC[3]  | 95.8   | 88.1   | 86.1 | 91.4   | 95.5   | 74.9 |
| IANet[4]   | 94.4   | 83.1   | 87.1 | -      | 73.4   |     |
| MMGA [5]   | 95.0   | 87.2   | 89.5 | -      | 78.1   |     |
| MHN-6(PCB) [6] | 95.1 | 98.1   | 85.0 | 89.1   | 94.6   | 77.2 |
| ABD[7]     | 95.6   | 88.3   | 89.0 | -      | 78.6   |     |
| HA-CNN[8]  | 91.2   | 75.7   | 80.5 | -      | 63.8   |     |
| reid-strong [13] | 94.5 | 85.9   | 86.4 | -      | 76.4   |     |
| Pyramid [19] | 95.7 | 98.4   | 88.2 | 89.0   | 94.7   | 79.0 |
| SCSN[20]   | 95.7   | 88.5   | 90.1 | -      | 79.0   |     |
| MGN[21]    | 95.7   | 86.9   | 88.7 | -      | 78.4   |     |
| PCB[22]    | 93.8   | 97.5   | 81.6 | 83.3   | 90.5   | 69.2 |
| SCPNet[23] | 91.2   | 75.2   | 80.3 | -      | 62.6   |     |
| MultiScale[24] | 88.9 | 73.1   | 79.2 | -      | 60.6   |     |
| AACN[25]   | 85.9   | 66.9   | 76.8 | -      | 59.3   |     |
| SVDNet[26] | 82.3   | 92.3   | 62.1 | 76.7   | 86.4   | 56.8 |

3.4. **Ablation Study**

In this section, in order to verify the effectiveness of the Pyramid Joint Attention(PJA) and Similarity Feature Fusion Branch(SFF) in PS-Net, we conduct multi-le incremental experiments on different components in the Market-1501 dataset. It is worth noting that in order to eliminate the influence of other factors, other irrelevant settings are the same in each ablation experiment. The baseline model we used is the ID-discriminative Embedding[IDE][18] network. In table 2, we verify the effectiveness of our proposed PJA and SFF. Note that the ‘2048’ and ‘512’ denotes the output feature dimension.

**Effect of SFF:** According to the table, we can see that our proposed SFF can effectively improve the performance of the network and generate more discriminating features. The rank-1 and mAP accuracy can reached 95.3% and 89.6% respectively.

**Effect of PJA:** Pyramid Joint Attention(PJA) take full advantage of the two different attention mechanisms and learn more context information. The ablation study shows that the PJA can effectively capture more context information and increase the rank-1 and mAP accuracy by 0.4% and
4.1% respectively.

| Method        | Rank-1 | mAP  |
|---------------|--------|------|
| IDE$^{2048}$ (baseline) | 94.9   | 85.5 |
| IDE$^{512}$ + SFF      | 95.3   | 89.6 |
| IDE$^{512}$ + PJA      | 95.3   | 88.8 |
| PS-Net          | 95.9   | 89.7 |

### 4. Conclusion

In this paper, we propose a new end-to-end network named Pyramid and Similarity Based Feature Enhancement Network (PS-Net) which includes Pyramid Joint Attention (PJA) and Similarity Feature Fusion Branch (SFF). Among them, the PJA module captures more global context and detail information, alleviates the mismatch of information obtained by pedestrians occupying different proportions in data sample by combines Pyramid Spatial Attention (PSA) and Channel Attention (CA). The SFF fuses the features which have different levels of information but have congruent relationships, enriches the information contained in pedestrian features and generates more robust feature. Extensive experiments on two different datasets shows that PS-Net can effectively improve the Re-ID performance and surpass most state-of-the-art methods. The ablation experiments shows that each component has a great contribution to its final performance.

### Acknowledgments
This work is supported in part by the National Natural Science Foundation of China under Grant 61771180 and Grant 61876056. Thank you very much for encouragement.

### References

[1] Li W, Zhao R, Xiao T, et al. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification[C]. computer vision and pattern recognition, 2014, pp. 152-159.

[2] Yi D, Lei Z, Liao S, et al. Deep Metric Learning for Person Re-identification[C]. international conference on pattern recognition, 2014, pp. 34-39.

[3] Zhang Z, Lan C, Zeng W, et al. Relation-Aware Global Attention for Person Re-Identification[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[4] Hou R, Ma B, Chang H, et al. Interaction-And-Aggregation Network for Person Re-Identification[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[5] Cai H, Wang Z, Cheng J, et al. Multi-Scale Body-Part Mask Guided Attention for Person Re-Identification[C]. computer vision and pattern recognition, 2019.

[6] Chen B, Deng W, Hu J, et al. Mixed High-Order Attention Network for Person Re-Identification[C]. international conference on computer vision, 2019, pp. 371-381.

[7] Chen T, Ding S, Xie J, et al. ABD-Net: Attentive but Diverse Person Re-Identification[C]. international conference on computer vision, 2019, pp. 8351-8361.

[8] Li W, Zhu X, Gong S, et al. Harmonious Attention Network for Person Re-identification[C]. computer vision and pattern recognition, 2018, pp. 2285-2294.

[9] Zhao H, Tian M, Sun S, et al. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion[C]. computer vision and pattern recognition, 2017, pp. 907-915.

[10] Lin T, Dollar P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]. computer vision and pattern recognition, 2017, pp. 936-944.

[11] Ghiassi G, Lin T, Le Q V, et al. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection[C]. computer vision and pattern recognition, 2019, pp. 7036-7045.
[12] Hu J, Shen L, Albanie S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, pp. 1-1.
[13] Luo H, Gu Y, Liao X, et al. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification[C]. computer vision and pattern recognition, 2019.
[14] Hermans A, Beyer L, Leibe B, et al. In Defense of the Triplet Loss for Person Re-Identification[J]. arXiv: Computer Vision and Pattern Recognition, 2017.
[15] Wen Y, Zhang K, Li Z, et al. A Discriminative Feature Learning Approach for Deep Face Recognition[C]. european conference on computer vision, 2016, pp. 499-515.
[16] Zheng L, Shen L, Tian L, et al. Scalable Person Re-identification: A Benchmark[C]. international conference on computer vision, 2015, pp. 1116-1124.
[17] Ristani E, Solera F, Zou R S, et al. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking[C]. european conference on computer vision, 2016, pp. 17-35.
[18] Zheng Z, Zheng L, Yang Y, et al. A Discriminatively Learned CNN Embedding for Person Reidentification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2017, 14, (1).
[19] Zheng F, Deng C, Sun X, et al. Pyramidal Person Re-IDentification via Multi-Loss Dynamic Training[C]. computer vision and pattern recognition, 2019, pp. 8514-8522.
[20] Chen X, Fu C, Zhao Y, et al. Salience-Guided Cascaded Suppression Network for Person Re-Identification[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[21] Wang G, Yuan Y, Chen X, et al. Learning Discriminative Features with Multiple Granularities for Person Re-Identification[C]. acm multimedia, 2018, pp. 274-282.
[22] Sun Y, Zheng L, Yang Y, et al. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)[C]. european conference on computer vision, 2018, pp. 501-518.
[23] Fan X, Luo H, Zhang X, et al. SCPNet: Spatial-Channel Parallelism Network for Joint Holistic and Partial Person Re-Identification[J]. arXiv: Computer Vision and Pattern Recognition, 2018.
[24] Chen Y, Zhu X, Gong S, et al. Person Re-identification by Deep Learning Multi-scale Representations[C]. international conference on computer vision, 2017, pp. 2590-2600.
[25] Xu J, Zhao R, Zhu F, et al. Attention-Aware Compositional Network for Person Re-identification[C]. computer vision and pattern recognition, 2018, pp. 2119-2128.
[26] Psichogios D C, Ungar L H. SVD-NET: an algorithm that automatically selects network structure[J]. IEEE Transactions on Neural Networks, 1994, 5, (3), pp. 513-515.