Complex & Intelligent Systems (2022) 8:5329–5342
https://doi.org/10.1007/s40747-022-00699-5

ORIGINAL ARTICLE

Gradient-supervised person re-identification based on dense feature pyramid network

Shaoqi Hou[1] · Kangning Yin[2] · Jie Liang[1] · Zhiguo Wang[2] · Yixi Pan[3] · Guangqiang Yin[2]

Received: 16 May 2021 / Accepted: 19 October 2021 / Published online: 19 May 2022

Abstract

In the monitoring scene, parameters of different cameras are vary greatly, which makes Person re-identification (Re-ID) tasks extremely susceptible to factors such as scale, blur, and occlusion. To alleviate the these problems, this paper proposes a Dense Feature Pyramid Network (DFPN), which can converge to a better performance without pretraining. To be more specific, DFPN is composed of three main parts. First, a new Residual Convolutional Block (RCB) is designed by referring to the construction method of ResBlock. Taking RCB as a basic unit and combining it with the convolution layer structure of VGGNet, we construct the backbone RVNet (Residual VGGNet) to realize the rapid convergence of the network and solve the disappearance of the gradient. Second, based on Feature Pyramid Network, we design the Dense Pyramid Fusion Module by integrating the connection mode of DenseNet, which aims at the improvement of the richness and scale diversity of feature maps by taking semantic information and detail information into account. Finally, to increase the receptive field of the feature map, we introduce an improved retinal receptive field structure Improved RFB (IRFB) on the basis of Receptive Field Block (RFB), which can effectively solve the problem of pedestrian occlusion. In experiments on the public datasets Market1501, DukeMTMC-reID and Occluded-Duke, the Rank-1 accuracy can reach 94.12%, 87.25% and 51.72% with pretraining, respectively. A series of ablation experiments and comparative experiments have proved the effectiveness of our modules and overall scheme.

Keywords Person re-identification · Residual connection · Feature pyramid · Receptive field

Introduction

Re-ID is based on the whole-body information of pedestrians. It aims to use computer vision technology to deal with the matching problem of the same pedestrian at different time and different locations, and then compensate for the visual limitations of fixed cameras. As an important subtask in the field of image retrieval, Re-ID makes up for the defect that pedestrian identity is difficult to confirm due to the lack of face information in surveillance video, and has broad application prospects in community, supermarket, airport and other scenes.

The concept of "Person Re-identification" was first mentioned by Zajdel et al. [1], from the University of Amsterdam, in a 2005 article on multi-camera target tracking. Due to the complex extraction process and poor representation ability of manual feature [2,3], the traditional Re-ID algorithms can only deal with the simple problems of pedestrian appearance and perspective difference under cross-camera. Until 2014, influenced by the strong learning ability of Convolutional

✉ Guangqiang Yin
yingq@uestc.edu.cn

Shaoqi Hou
sqhou@std.uestc.edu.cn

Kangning Yin
knyin@std.uestc.edu.cn

Jie Liang
liangjie@std.uestc.edu.cn

Zhiguo Wang
zgwang@uestc.edu.cn

Yixi Pan
yxpan@std.uestc.edu.cn

[1] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

[2] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

[3] Glasgow College, University of Electronic Science and Technology of China, Chengdu 611731, China

Neural Network (CNN), deep learning in image classification began to be applied to the Re-ID field [4–7]. With the rapid development of computer vision technology, research on Re-ID based on deep learning has sprung up since 2016.

The most significant difficulty of Re-ID based on deep learning is that, in the cross-camera monitoring images, the whole-body image changes greatly in scale, posture and resolution, which makes the characteristics of the same pedestrian vary greatly. In addition, the inevitable occlusion environment leads to the lack of pedestrian saliency information, which further affects the accuracy of Re-ID. Therefore, the most serious problems of scale, blur, occlusion and other issues under cross-camera have always been the focus of many researchers.

In response to these problems, the current mainstream Re-ID algorithms are mainly divided into two categories: representation learning and metric learning. Representation learning mainly extracts or combines more robust features to strengthen the discrimination of pedestrian features [8]. The famous attention mechanism is an outstanding representative in this regard. Because of the good effect on the salient feature extraction of pedestrians, this mechanism has been widely used in the field of Re-ID in recent years [9–11]. Meanwhile, other scholars mainly focus on the metric learning, and are committed to designing better metric functions, so that the intra-class distance of pedestrian features is smaller and the inter-class distance is larger. Triplet loss function [12], as one of the most basic and attractive Distance Losses, has derived many valuable metric functions. Of course, with the advantages of joint training becoming increasingly prominent, more and more researchers choose to integrate the idea of representation learning and metric learning. In addition to representation learning and metric learning, there have been some other Re-ID methods based on local features in recent years. The main idea is to align and integrate pedestrian local features through image segmentation [13], skeleton key point positioning [14] and posture correction [15], and finally improve the discrimination of Re-ID tasks using local features as supplementary information.

Although the above typical methods can greatly improve the performance of Re-ID, they bring many new problems at the same time: attention mechanism can enhance pedestrian salient feature extraction, but it will introduce a large number of new parameters, which greatly increases the computational complexity of the model. The design of loss function in metric learning is complex and difficult to implement. Although the Re-ID methods based on local feature are simple, they bring a lot of redundant computation and affect the speed of Re-ID.

Different from the above classical algorithms, we propose a new framework DFPN based on dense feature pyramid from the direction of representation learning, which has high flexibility because it can be trained from scratch. In particular, our DFPN transfers the semantic information of the high-level feature map downward step by step, and makes it fully integrated with the details of the low-level feature map. Without introducing external feature enhancement modules, DFPN improves the richness, balance and integrity of pedestrian information in cross-camera scenes, and then improves the representation ability of pedestrian feature vectors. The specific innovations are as follows:

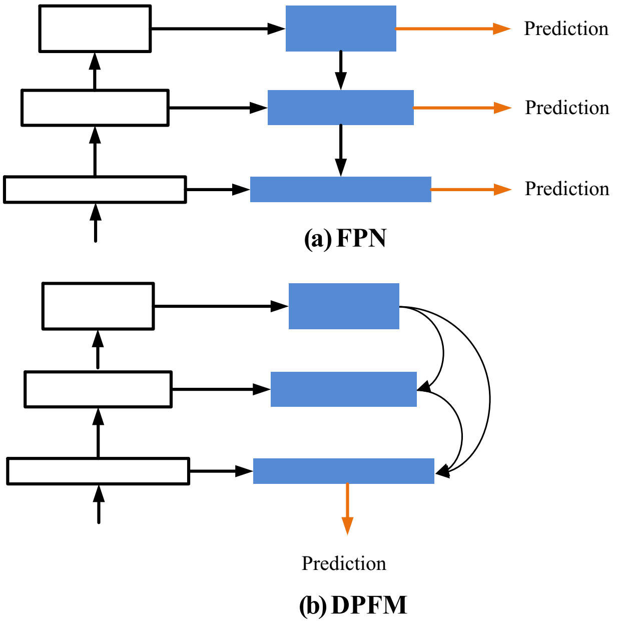

Fig. 1 The structure comparison of FPN and our DPFM. **a** FPN, **b** DPFM

1. Feature extraction: we build backbone RVNet, which has a simple structure, fewer parameters and easy training.
2. Feature enhancement module I: we design Dense Pyramid Fusion Module (DPFM) module (as shown in Fig. 1b), which combines high-level semantic information and low-level detail information, and improves the scale invariance of feature map.
3. Feature enhancement module II: we introduce Improved RFB (IRFB) to solve the separation of context information of feature map and effectively deal with the pedestrian matching problem under blur and occlusion conditions.

The remainder about the paper is organized as follows: The next section introduces some related works. The subsequent section demonstrates the presented recommendation framework and modules. Then penultimate section shows the design of experiments and analysis about results. The final section illustrates the conclusion.

Related works

Feature extraction in Re-ID

The key step of computer vision tasks based on deep learning, such as object detection, Re-ID, is to select the appropriate backbone. The depth and structure design of backbone largely affect the final performance of the network. In 2012, AlexNet [16] won the championship in the ImageNet competition and became the first work of convolutional neural network to solve large-scale dataset classification tasks. VGGNet [17] is the earliest deep convolutional neural network after the rise of deep learning, which proves that increasing the depth of the network can significantly improve the performance of the network. Compared with AlexNet, VGGNet uses continuous 3*3 convolution layers instead of larger convolution layers in AlexNet (such as 11*11 and 7*7), which has a lower computational cost. At the same time, multiple nonlinear layers can increase the network depth to ensure that more complex patterns are learned.

Although model performance can be improved by deepening network structure, deeper networks often have more parameters. In addition, since the optimization of the network needs the back propagation algorithm, as the number of layers increases, the gradient of the network is prone to disappear or explode, which makes the model difficult to train and converge. Aiming at the problem of deep network degradation, He et al. proposed ResNet [18] in 2015, which shows a convenient way to build deeper neural networks with ResBlock. Later DenseNet [19] also pointed out that if the layer close to the input and the layer close to the output has a shorter connection, it is easy to construct and train a deeper network, and thus effectively improving the utilization and transmission efficiency of features. For example, Highway Networks [20] and Stochastic Depth [21] both provided a shorter connection between the front layer and the back layer.

Feature enhancement in Re-ID

In the field of Re-ID based on deep learning, the main ways of feature enhancement are through attention mechanism and feature fusion. Attention mechanism can focus on the region of interest to enhance salient features, so it is widely used in the field of Re-ID and has achieved good results [22–24]. However, occlusion will cause the lack of key information, which will greatly reduce the performance of these methods. As another means of feature enhancement, feature fusion can compensate the shortcoming of attention mechanism.

As a typical representative of feature fusion methods, Feature Pyramid Network (FPN) [25] solves the problem of variable object scale (as shown in Fig. 1a). Before this, most detectors based on deep learning only detect at the top feature map of the network. FPN uses a top-down architecture with horizontal connections to construct high-level semantics in feature maps at all levels. In addition, DenseNet adopts dense connection, which strengthens the reuse of features and makes the network easier to train. Although FPN can solve multi-scale problems well, it shows no effective improvement in dealing with blur and occlusion of object images. Increasing the receptive field is a key point to solve the problem of object occlusion. Liu [26] introduced RFB in SSD network to strengthen the feature extraction ability of the network by simulating the receptive field of human vision. RFB draws on the multi-branch fusion idea of Inception [27] in structure. The main difference is that dilated convolution is added in RFB, which effectively increases the receptive field. In the field of Re-ID, using abstract high-level features or single feature map cannot solve the problems of multi-scale and occlusion of pedestrian objects well, so it is an innovation to use appropriate feature fusion methods to combine deep and shallow features.

Proposed methods

This section consists of four subsections. First, we introduce the overall framework of the proposed DFPN in the first subsection; then, in the second to the fourth subsections, we successively elaborate the design scheme of several important modules including RVNet, DPFM and IRFB.

Introduction of the overall framework

As shown in Fig. 2, the DFPN proposed in this paper mainly includes four modules: feature extraction module RVNet, feature enhancement module I DPFM, feature enhancement module II IRFB and feature aggregation module Heads. The first three of them are our newly designed modules, and perform different functions in the whole architecture:

1. RVNet, as the backbone of our network, is one of the most important modules in DFPN. RVNet consists of five blocks from Block1 to Block5, and each block is composed of several designed RCBs. RCB can not only effectively control the number of channels in the feature map, but also effectively alleviate the network degradation problem by adopting the design idea of "Shortcut" [20].
2. To solve the multi-scale problem of the object, we put the extracted features into DPFM, whose essence is feature fusion. DPFM selects multiple high-level feature maps in RVNet, recursively transmits the rich semantic information in the high-level feature map to the bottom by dense connection, and fuses them in a specific way with the low-level feature map. The fused feature map

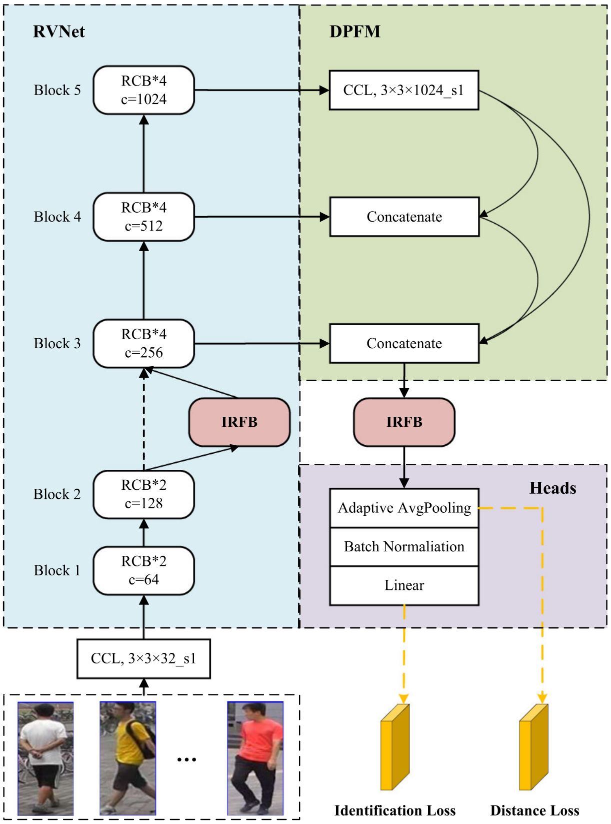

Fig. 2 Overall framework of our DFPN

takes into account global semantic information and local detail information, which can enhance the constraint and expression of pedestrian salient features.

3. In view of the occlusion and ambiguity of the object, we design the IRFB and insert it into the specific location of the network. IRFB is improved on the basis of RFB [26]. Unlike RFB, we equivalently replace the convolution layer of larger kernels in RFB, reducing parameters while adding more nonlinear mappings. The IRFB can assist the network to effectively extract the context feature information of the same pedestrian object and improve the discriminant ability of the model for occlusion objects.
4. The final feature aggregation module Heads is composed of Adaptive AvePooling layer, batch normalization (BN) layer and fully connected layer (i.e., Linear layer). The purpose is to integrate the highly abstract features after multiple convolutions, and provide interfaces for Distance Loss and Identification Loss. In particular, the pooling layer and the normalization layer can reduce the dimension of features and obtain stable data distribution respectively.

Specifically, combined with Fig. 2, the algorithm flow of our DFPN is as follows:

Algorithm 1 flow: DFPN

```
Input: Img ( N*C*H*W )
  Step 1: Feature map conversion: Img is converted into feature map F0 (N*32*H0*W0) through a CCL layer;
  Step 2: Feature extraction: input the feature map F0 to RVNet, and get the feature maps F1 ∼ F5 (N*32*H1*W1 ∼ N*1024*H5*W5) by Block1 ∼ 5;
  Step 3: Feature enhancement processing one: select the feature maps F3 ∼ F5 to construct the DPFM, and feature map F6 (N*512*H6*W6) is obtained after the original image undergoes DPFM;
  Step 4: Feature enhancement processing two: input feature map F6 into the IRFB to obtain feature map F7 (N*512*H7*W7);
  Step 5: Feature aggregation: input the feature map F7 into Heads and divide it into two branches after Adaptive AvePooling layer: branch one for metric learning to compute Distance Loss. Branch 2 continues to pass down, and then passes through BN layer and Linear layer to compute Identification Loss.
Output: Distance Loss, Identification Loss
```

Design of RVNet

The improvement of CNNs has made great progress in recent years, but the main networks represented by VGGNet and ResNet still have the following defects: the former one is simple in structure, usually uses smaller size convolution layers, and has fewer parameters but is not easy to converge. The latter one uses the idea of “Shortcut” to solve the problem of network degradation during training, but its deep network structure has increased the amount of calculation. In addition, both VGGNet series and ResNet series need pretraining on ImageNet large dataset before migrating to specific tasks, which makes the models less flexible. To this end, we design a new backbone RVNet, which combines the simple structure of VGGNet with the advantages of “Shortcut”. It not only greatly reduces parameters, but also enables the network to train from scratch.

Figure 2 shows that the RVNet is composed of five blocks with RCB as the basic unit according to the simplified structure of VGGNet-19 (see Table 1 for network structure). The essence of RVNet is the designed RCB block. Figure 3b shows that RCB is composed of two composite convolutional layers (CCLs) in the form of “Conv-BN-LeakyReLU”. The difference between the two CCLs is that the first CCL uses a 1*1 convolution layer and the second uses a 3*3 convolution layer, but both have a stride of 1. First, the convolution layer of 1*1 can compress the number of channels of the feature map without changing the size of the feature map to reduce subsequent computation. Second, the convolution layer of 3*3 is used to restore the number of channels of the feature map and process the feature information. Since the convolution layer of 3*3 has fewer parameters, the calculation of the network is further reduced. In particular, through the operation of dimension reduction and dimension increase of the number of channels on the input feature map, the interaction of cross-channel information can be increased, and then a more representative feature vector can be generated.

At the same time, referring to the design concept of ResBlock, we add "Shortcut" branch in each RCB to supervise the shallow feature mapping and solve the problem that it is difficult to train due to the gradient divergence in the deep network. From Fig. 3, it should be noted that our RCB has obvious differences compared with ResBlock, which are listed as follows:

1. The activation function used in RCB is LeakyReLU, while the ReLU used by ResBlock. As a variant of ReLU, LeakyReLU also gives a small gradient to the negative value, which is convenient to retain the integrity and diversity of feature information.
2. RCB alternately uses the convolution layers of 1*1 and 3*3, reduces parameters and improves the nonlinearity of the network by reducing and increasing the dimension of the channel number of the feature map, while the two 3*3 convolution layers used by ResBlock keep the number of channels constant.
3. The feature fusion operation used by RCB is "Concat", while ResBlock uses "Element-wise add" operation. The "Concat" operation can better aggregate multi-level features and will not cause confusion of feature information as "Element-wise add".

Table 1 The structure comparison of VGGNet-19 and RVNet

Number	VGG19	RVNet
1	Conv3-64	RCB-64 (32, 64)
2	Conv3-64	RCB-64 (32, 64)
	Maxpool	
3	Conv3-128	RCB-128 (64, 128)
4	Conv3-128	RCB-128 (64, 128)
	Maxpool	
5	Conv3-256	RCB-256 (128, 256)
6	Conv3-256	RCB-256 (128, 256)
7	Conv3-256	RCB-256 (128, 256)
8	Conv3-256	RCB-256 (128, 256)
	Maxpool	
9	Conv3-512	RCB-512 (256, 512)
10	Conv3-512	RCB-512 (256, 512)
11	Conv3-512	RCB-512 (256, 512)
12	Conv3-512	RCB-512 (256, 512)
	Maxpool	
13	Conv3-512	RCB-1024 (512, 1024)
14	Conv3-512	RCB-1024 (512, 1024)
15	Conv3-512	RCB-1024 (512, 1024)
16	Conv3-512	RCB-1024 (512, 1024)

The two numbers in parentheses are the channel numbers of Conv1 and Conv3 respectively

Design of DPFM

In the field of computer vision, FPN is an effective method to solve the multi-scale problem of the object. It constructs the feature pyramid on the multi-layer feature map in the form of top-down, and then uses the feature maps of different scales on the pyramid to predict. However, there are still many problems in the construction of FPN:

1. High-level feature map has strong discriminative semantic information, while FPN only uses adjacent upper-level feature map information when constructing, which leads to a low utilization of other high-level feature information.
2. The "Element-wise add" feature fusion method used by FPN destroys the hierarchical information of the feature map, resulting in feature confusion.
3. Multi-scale prediction can make the results more accurate, but it also increases the amount of calculation.

To solve these problems, we use the structural advantages of DenseNet to propose a Dense Pyramid Fusion Module (DPFM). Different from FPN, we apply dense connection to DPFM, recursively transmit the rich semantic information of the high-level feature map to the bottom, and adopt a specific way (the ratio of the number of channels of the current layer feature map to the sum of channels of all downwardly transmitted high-level feature maps = 1:1) to concatenate with the low-level feature map. Finally, only the end feature map is used for prediction, as shown in Fig. 4. Compared with FPN, DPFM has significant advantages: dense connection makes the loss function in the network provide supervision signal to not only the final output layer, but also the front non-output layer in the network (i.e., all high-level feature maps recursively transmitted), which is not only conducive to the deep mining of high-level semantic information, but also helpful in alleviating the degradation problem in network training. On the other hand, compared with the direct "Element-wise add", the channel concatenation effectively retains the integrity of the hierarchical relationship of aggregated features, and takes into account global semantic information and local detail information, thus effectively responding to the appearance changes such as pedestrian scale under different devices.

The specific construction process of DPFM is as follows: firstly, the feature map C5 output by Block5 is processed by CCL to generate Feature5 (as shown in Eq. (1), and the size and number of channels of C5 remain unchanged during the processing; secondly, deep feature map Feature5, which contains rich semantic information, generates Feature4 (as shown in Eq. (2)) by manipulating Op1 (as shown in Fig. 5a)

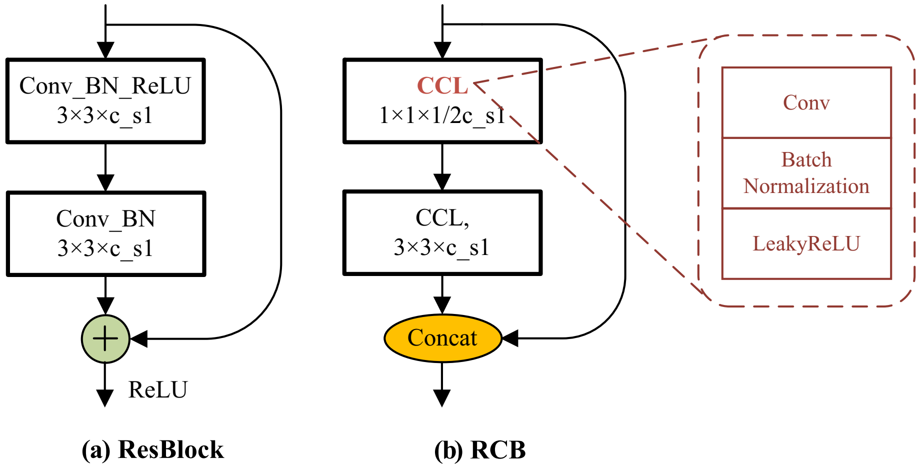

Fig. 3 The structure comparison of RCB and ResBlock. **a** ResBlock, **b** RCB

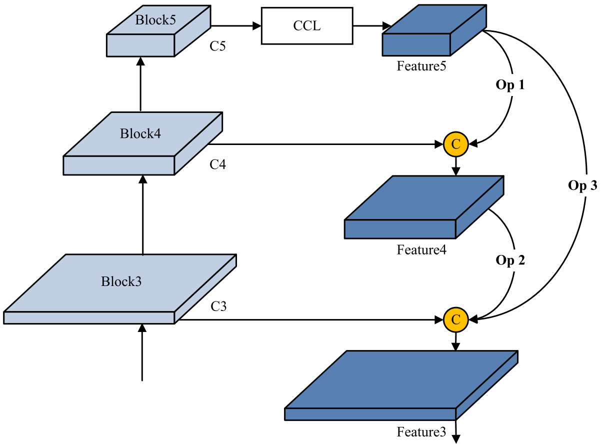

Fig. 4 The structure of DPFM

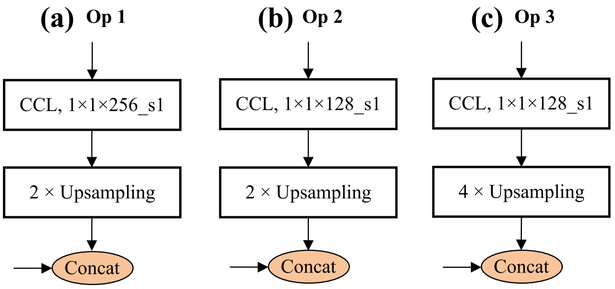

Fig. 5 The specific operations for Op1, Op2 and Op3. **a** Op1, **b** Op2, **c** Op3

and concatenating it with Block4's output feature map C4 in the form of channel ratio 1:1; then, Feature5 and Feature4 are concatenated with the feature map C3 output by Block3 through Op2 (as shown in Fig. 5b) and Op3 (as shown in Fig. 5c), respectively, to generate the feature map Feature3 (as shown in Eq. (3)); finally, the end feature map combined with shallow detail features and deep semantic features is output to discriminate the Re-ID task.

$$\text{Feature5} = \text{CCL}(\text{C5}) \tag{1}$$

$$\text{Feature4} = [\text{C4}, \text{Op1}(\text{Feature5})] \tag{2}$$

$$\text{Feature3} = [\text{C3}, \text{Op2}(\text{Feature4}), \text{Op3}(\text{Feature5})]. \tag{3}$$

Design of IRFB

When the object is occluded, the surrounding information can often be used to assist the discrimination. Inspired by this, RFB effectively increases the receptive field of the feature map by introducing dilated convolution, thus introducing more context-assisted information. At the same time, RFB further improves the scale invariance of the model using multi-branch structure, as shown in Fig. 6a. Based on RFB, we propose IRFB. The specific improvement of structure is as follows: as shown in Fig. 6b, the 5*5 convolution layer in RFB is replaced by two serial 3*3 convolution layers with "stride=1" and "padding=1". In particular, when passing through the first 3*3 convolution layer, the number of channels in the feature map is compressed to half of the original, and then after passing through the second 3*3 convolution layer, the number of channels in the feature map is restored to the corresponding value. Simply put, we achieve the compression and recovery of the feature map's channels by controlling the number of the two 3*3 convolutional layers and the number of their channels. In view of its function, we insert IRFB into the intersection of Block3 and Block4 in RVNet, and after the DPFM module, to enhance the sensitivity of IRFB module to the scale change of feature map. In addition, we also introduced the "Shortcut" technique in IRFB, which makes it more conducive to backpropagate the gradient during training.

Compared with RFB, IRFB has the following advantages:

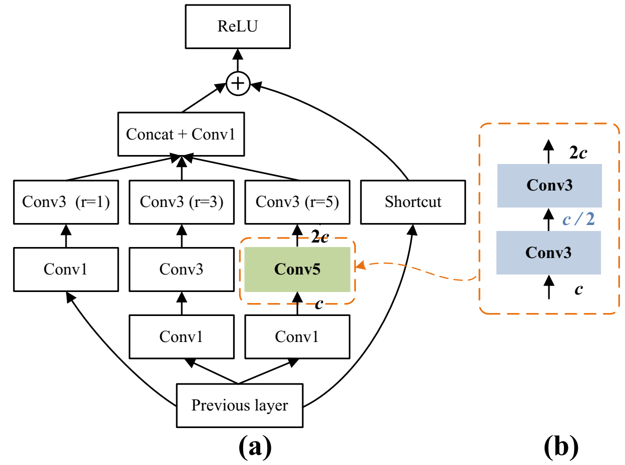

Fig. 6 The structure diagram before and after improvement. **a** RFB, **b** IRFB

1. Due to the mapping of two activation functions, IRFB has more nonlinearity.
2. IRFB has fewer parameters while keeping the size of the receptive field unchanged.

The calculation formula of receptive field R is

$$R_k = R_{k-1} + \left[(f_k - 1) \times \prod_{i=1}^{k-1} s_i \right]. \tag{4}$$

Among them, R_{k-1} is the receptive field size corresponding to the $(k-1)$th layer, f_k is the convolution kernel size of the kth layer, and s_i is the stride of the ith layer. According to Eq. (4), the size of the receptive field before and after the improvement is calculated as follows, the intuitive schematic diagram is shown in Fig. 7:

$$\begin{aligned} R_{\text{RFB}} &= 5 \\ R_{\text{IRFB}} &= 3 + (3-1) \times 1 = 5. \end{aligned} \tag{5}$$

The number of parameters of the two modules is calculated as follows:

$$\begin{aligned} S_{\text{RFB}} &= 5 \times 5 \times c \times 2c = 50c^2 \\ S_{\text{IRFB}} &= 3 \times 3 \times c \times \frac{1}{2}c + 3 \times 3 \times \frac{1}{2}c \times 2c = \frac{27}{2}c^2. \end{aligned} \tag{6}$$

From the Eq. (6) it can be seen that the parameters of IRFB are reduced to about one-fourth of the RFB.

Experiments and analysis

This section contains four subsections. The specific arrangement is as follows: first, we introduce the datasets and evaluation indicators used; second, we elaborate on some specific experimental details and parameter selection; then in the third subsection, we set up ablation experiments to verify the effectiveness of each design; finally, in the fourth and fifth subsections, we show the comparisons with classical and advanced methods respectively, highlighting the superiority of our overall model. It should be noted that all our experiments are trained and tested on a single dataset.

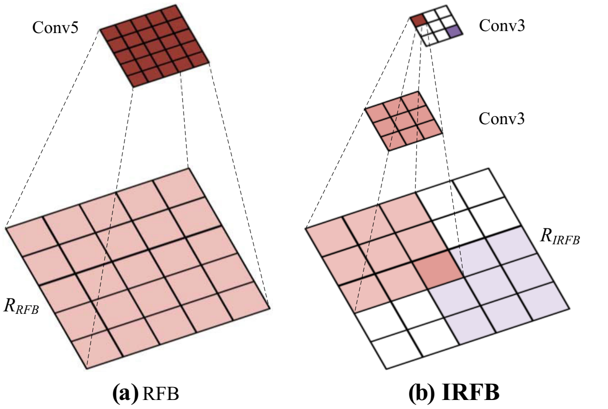

Fig. 7 The Receptive field diagram before and after improvement. **a** RFB, **b** IRFB

Datasets and evaluation indicators

Datasets

Market1501 [28]: Market1501 contains 1501 pedestrian IDs and 32,668 detection images captured by 6 cameras. Each pedestrian is captured by at least 2 cameras, and there may be multiple images under one camera. Among them, the training set has 751 pedestrian IDs and contains 12,936 images; the query set has 750 pedestrian IDs and contains 3368 images; the test set has 16,384 images with 750 pedestrian IDs.

DukeMTMC-reID [29]: DukeMTMC is a large-scale multi-target multi-camera pedestrian tracking dataset DukeMTMC-reID is a person re-identification subset of DukeMTMC, and provides a manually labeled bounding box. DukeMTMC-reID contains 16,522 training images of 702 pedestrian IDs, 2228 query images from another 702 pedestrian IDs, and a search gallery of 17,661 images containing 702 pedestrian IDs (the same as the pedestrian IDs of the query images).

Evaluation indicators

We adopt the cumulative matching characteristics (CMC) at Rank-1 and the mean average precision (mAP) as the evaluation indicators to test the performance of different Re-ID methods on these datasets. The mAP is the mean of average precision (AP) for each query image. Rank-1 is the probability that the top image in the search results is the object.

Implementation details and loss function

Implementation details

To ensure the consistency of the experimental results, the experimental process is carried out in the same software and hardware environment. We use the fast-reid framework [30] to build the model, the output feature dimension is set to 512 by default, and all designed models are not pretrained. The computing platform uses single GPU (memory 11 GB) of GTX 1080Ti, pytorch uses version 1.7.0, CUDA's version is 11.0 and python's version is 3.8.3. In the training phase, we scale all images to a size of 256*128 and perform random erasure operations with a probability of 0.5. In each batch, we randomly select 16 pedestrian IDs, and randomly sample 4 images for each ID. The Adam optimizer is used in the training process to update the model parameters. The initial learning rate is set to 1e−4, and the learning rate uses a multistep strategy. When iterating to 120, 270, and 360 epochs, the learning rate is reduced to 0.1 times the original, and the training is stopped after a total of 390 epochs.

Loss function

The Triplet Loss [31] was selected for our Distance Loss, and its formula calculated as follows:

$$L_{\text{tri}} = [d_{\text{p}} - d_{\text{n}} + \alpha]. \tag{7}$$

Among them, d_{p} is the Euclidean distance between the target and the positive samples, d_{n} is the Euclidean distance between the target and the negative samples, and α is the threshold. The goal of Triplet Loss is to make d_{p} smaller and make d_{n} larger through optimization.

For the Identification Loss, we chooses Softmax Loss which can map the output values of multiple neurons to (0, 1) interval. Suppose there are n samples in a batch, and the calculation formula of Softmax Loss is:

$$L_{\text{soft}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_{ij})\log(q(x_{ij})). \tag{8}$$

Among them, m represents the number of classes, $p(x_{ij})$ and $q(x_{ij})$, respectively, represent the predicted probability and the true probability of the target sample i belonging to class j.

Table 2 Performance comparison of different activation functions in CCL

Methods	Rank-1 (%)	mAP (%)
VGGNet-19 + RCB (ReLU)	88.27	72.24
VGGNet-19 + RCB (PReLU)	85.93	68.48
VGGNet-19 + RCB (LeakyReLU)	88.66	72.30

Ablation experiments

To ensure that there is only one independent variable for each experiment, it is important to note that, except for the last subsection, the dataset we use in other subsections is Market1501, and if no special instructions is given, all experiments exclude the pretraining of models.

Analyze the performance of RVNet

To fully prove the effectiveness of RVNet, we use progressive experimental methods to verify the designed CCL layer and RCB block.

(1) The effectiveness of LeakyReLU in CCL.
Based on the most commonly used ReLU activation function, we select the two most representative variants PReLU [32] and LeakyReLU [33] for comparative experiments. In the experiment, the order of each basic layer in CCL is "Conv-BN-Activ".
As shown in Table 2, ReLU can achieve 88.27% Rank-1 and 72.24% mAP on Market1501 dataset. After using the PReLU, compared with using the ReLU, Rank-1 decreases by 2.34% and mAP decreases by 3.76%. After using LeakyReLU, compared with using ReLU, Rank-1 increases to 88.66% (increases by 0.39%), mAP increases to 72.30% (increases by 0.06%).
In summary, with the use of LeakyReLU in CCL, it gives the best result, indicating that it is not desirable for ReLU to set all negative weights to 0, because it destroys the integrity of feature information. Without special instructions, we will use LeakyReLU by default in subsequent experiments.

(2) The effectiveness of different combination in CCL.
In this experiment, we verify the best combination of "Convolution", "Batch Normalization" and "LeakyReLU". Among them, "CBL" indicates that the order of the common layers in the CCL is "Conv-BN-LeakyReLU", other equivalents.
It can be seen from Table 3 that the accuracy of other combinations on Market1501 decreases to varying degrees without using the "CBL". Therefore, the combination of "CBL" is the best, which proves that the BN layer is generally used behind the convolution layer, before

Table 3 Performance comparison of different combination of CCL

Methods	Rank-1 (%)	mAP (%)
VGGNet-19 + RCB (LBC)	84.23	65.96
VGGNet-19 + RCB (LCB)	85.15	67.22
VGGNet-19 + RCB (BCL)	86.05	68.61
VGGNet-19 + RCB (BLC)	86.13	68.01
VGGNet-19 + RCB (CLB)	86.05	67.47
VGGNet-19 + RCB (CBL)	88.66	72.30

the nonlinear layer such as LeakyReLU: because the shape of the output distribution of the nonlinear layer changes during the training process, normalization cannot eliminate its variance offset. On the contrary, the output of convolution layer is generally a symmetric, non-sparse distribution (more similar to Gaussian distribution), which will produce a more stable distribution after normalization. If there is no special explanation, we will use "CBL" combination by default in the subsequent experiments.

It can be seen from Table 3 that the accuracy of other combinations on Market1501 decreases to varying degrees without using the "CBL". Therefore, the combination of "CBL" is the best one, which proves that the BN layer is generally used behind the convolution layer and before the nonlinear layer (such as LeakyReLU). We think the reason for above result is that the shape of the output distribution of the nonlinear layer changes during the training process, and normalization cannot eliminate its variance offset; on the contrary, the output of convolution layer is generally a symmetric, non-sparse distribution (more similar to Gaussian distribution), which will produce a more stable distribution after normalization. If there is no special explanation, we will use "CBL" combination by default in the subsequent experiments.

(3) The effectiveness of RCB.

It can be seen from Table 4 that the network using VGGNet-19 as backbone does not converge without pretraining. Using the pretrained model, VGGNet-19 can reach 66.63% Rank-1 and 40.27% mAP. However, compared with our RVNet (without pretraining), VGGNet-19 still has 22.03% Rank-1 gap and 32.03% mAP gap after pretraining. According to the loss curve of Fig. 8, our RVNet has faster convergence speed than VGGNet-19 (pretrained), and finally achieves a better result. This fully illustrates the design of RVNet can not only better feature extraction, but is also more favorable for gradient transfer in the training process.

Table 4 Performance comparison before and after using RCB

Methods	Rank-1 (%)	mAP (%)
VGGNet-19	0.15	0.15
VGGNet-19 (pretrained)	66.63	40.27
RVNet (VGGNet-19 + RCB)	88.66	72.30

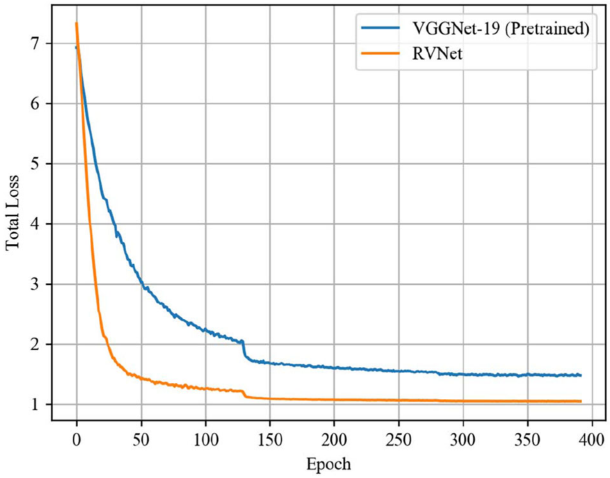

Fig. 8 Loss curve using different backbones

Analyze the performance of DPFM and IRFB

Based on RVNet, we further validate the effectiveness of the proposed DPFM and IRFB.

As shown in Table 5, the accuracy of the model is significantly improved by adding DPFM on the basis of RVNet, and Rank-1 and mAP increase by 2.2% and 4.48%, respectively, which proves that the fusion of high-level semantic information and low-level detail information is indeed conducive to enhancing the discrimination of Re-ID tasks. After the introduction of IRFB, the Rank-1 of the model increases from 90.86 to 92.93% (increased by 2.07%), and the mAP increases from 76.78 to 79.42% (increased by 2.64%), which fully shows that IRFB can effectively extract the context feature information of the same pedestrian target and improve the ability of identifying the occlusion targets.

To further highlight the superiority of our DPFM and IRFB, we compare the effects of DPFM and FPN, IRFB and RFB respectively. As shown in Table 6, from the first two rows of results, it can be seen that DPFM is 1.89% and 3.03% higher than FPN's Rank-1 and mAP, respectively, which proves that our DPFM better retains the structure of

Table 5 Performance comparison after using DPFM and IRFB

Methods	Rank-1 (%)	mAP (%)
VGGNet-19 (pretrained)	66.63	40.27
RVNet	88.66	72.30
RVNet + DPFM	90.86	76.78
RVNet + DPFM + IRFB (our DFPN)	92.93	79.42

Table 6 Performance comparison of FPN with DPFM and RFB with IRFB

Methods	Rank-1 (%)	mAP (%)
RVNet + FPN	88.97	73.75
RVNet + DPFM	90.86	76.78
RVNet + DPFM + RFB	91.89	77.30
RVNet + DPFM + IRFB	92.93	79.42

Table 7 Performance comparison of different feature dimensions

Methods	Rank-1 (%)	mAP (%)
DFPN (128)	88.63	74.85
DFPN (256)	91.06	76.90
DFPN (512)	92.93	79.42
DFPN (1024)	91.12	78.13
DFPN (2048)	90.62	78.55

hierarchical information. Similarly, comparing the experimental results in the third row and the fourth row, IRFB is also 1.04% and 2.12% higher than RFB on Rank-1 and mAP, respectively, indicating that information interaction between different channels is very important.

The experiments of the two groups have fully proved that our DPFM and IRFB are very effective, and also show that solving the multi-scale and occlusion problems of pedestrian targets is indeed helpful in the Re-ID task.

Analyze the performance of dimensions and datasets

It can be seen from Table 7 that the performance of the model comes best when the feature dimension is 512. Whether the feature dimension is larger or smaller, the performance of the model will decrease to varying degrees. To verify the generalization ability of our model, we test the model on another public dataset DukeMTMC-reID. From Table 8 below, our DFPN still has competitive performance in DukeMTMC-reID, Rank-1 and mAP can reach 83.89% and 69.58% respectively. In addition, DFPN shows a fast training and inference speed on both Market1501 and DukeMTMC-reID, especially reaching an inference speed of 887 FPS on Market1501.

Comparative experiment of classical algorithms

To prove the superiority of our DFPN in the overall design. We have selected some of the most representative algorithms in the Re-ID field for comparison, such as PAN [34], SVDNet [35], APRNet [36], REDA [37], DPFL [38], DuATM [39], PCB [13] and AANet [40]. The selection principles are as follows:

1. Their structures are based on convolutional neural networks;
2. They are representative algorithms in different Re-ID genres;
3. They both carried out experiments on the Market1501 and DukeMTMC-reID datasets, and both used the evaluation indicators Rank-1 and mAP.

The comparison results are shown in Table 9 below. It can be seen that the Rank-1 and mAP of our DFPN on Market1501 and DukeMTMC-reID datasets exceed most of classical algorithms. Among them, compared with PAN, DFPN has increased by more than 10% in various indicators on different datasets. However, compared with AANet, DFPN still has a gap of 1% and 3.76% in Market1501 and DukeMTMC-reID, respectively, which should be related to the simplicity of our backbone, because the backbone of AANet has more than 152 layers, while our RVNet has only a dozen layers.

In addition, we need to emphasize in particular that our DFPN is not pretrained on ImageNet [16] or COCO [41] large datasets because of our limited computing resources, and it is well known that pretrained models generally have a significant improvement in accuracy. Therefore, in general, the performance of our DFPN algorithm has underlying potential in future application.

Comparative experiment of advanced algorithms

In the previous experiments, we use the controlled variable method to verify that our DFPN without pretraining surpass most of the classic algorithms in and before 2019. To make our method more convincing, we have completed the pretraining of our DFPN on COCO dataset and have selected

Table 8 Performance of our model on different datasets

Methods	Rank-1 (%)	mAP (%)	Training speed (FPS)	Inference speed (FPS)
DFPN (Market1501)	92.93	79.42	127	887
DFPN (DukeMTMC-reID)	83.89	69.58	100	699

Table 9 Comparison with the classical algorithms on Market1501 and DukeMTMC-reID datasets

Algorithms	Market15011		DukeMTMC-reID	
	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
PAN [34]	82.20	63.40	71.59	51.51
SVDNet [35]	82.30	62.10	76.70	56.80
APRNet [36]	87.04	66.89	73.92	55.56
REDA [37]	87.08	71.31	79.31	62.44
DPFL [38]	90.00	70.60	77.60	58.60
DuATM [39]	91.40	76.60	81.80	64.60
PCB [13]	92.30	77.40	81.80	66.10
Our DFPN	**92.93**	**79.42**	**83.89**	**69.58**
AANet [40]	93.93	83.41	87.65	74.29

Bold represents our method

Table 10 Performance comparison of our DFPN before and after pretraining

Methods	Datasets	Pretrained	Rank-1 (%)	mAP (%)
DFPN	Market1501	No	92.93	79.42
		Yes	94.12	91.36
	DukeMTMC-reID	No	83.89	69.58
		Yes	87.25	84.37
	Occluded-Duke	No	43.35	34.95
		Yes	51.72	41.69
DFPN - IRFB	Occluded-Duke	Yes	43.53	35.09

several advanced algorithms proposed in the past 2 years to compare with our model.

To fully illustrate that our DFPN still has a good performance under occlusion condition, we specially introduce a new dataset named Occluded-Duke [42], which is by far the largest dataset for the occlusion person Re-ID. Occluded-Duke is selected from DukeMTMC-reID by leaving occluded images and filter out some overlap images. Its training set contains 15,618 images covering 702 identities and testing set contains 1110 identities in total, including 17,661 gallery images and 2210 query images.

Table 12 Comparison with the advanced algorithms on Occluded-Duke dataset

Algorithms	Rank-1 (%)	mAP (%)
HOReID 2020 [47] (SOTA)	55.10	43.80
PGFA 2019 [42]	51.40	37.30
Ad-Occluded 2018 [48]	44.50	32.20
SFR 2018 [49]	42.30	32.00
DSR 2018 [50]	40.80	30.40
Our DFPN (pretrained)	**51.72**	**41.69**

Bold represents our method

Table 11 Comparison with the advanced algorithms on Market1501 and DukeMTMC-reID datasets

Algorithms	Market15011		DukeMTMC-reID	
	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
FlipReID 2021 [43] (SOTA)	95.80	94.70	93.00	90.70
DeepMiner 2021 [44]	95.70	90.40	91.20	81.80
CircleLoss 2020 [45]	96.10	87.40	89.00	79.60
FastReID 2020 [30]	95.40	88.20	89.60	79.80
CBN 2020 [46]	91.30	77.30	82.50	67.30
Our DFPN (pretrained)	**94.12**	**91.36**	**87.25**	**84.37**

Bold represents our method

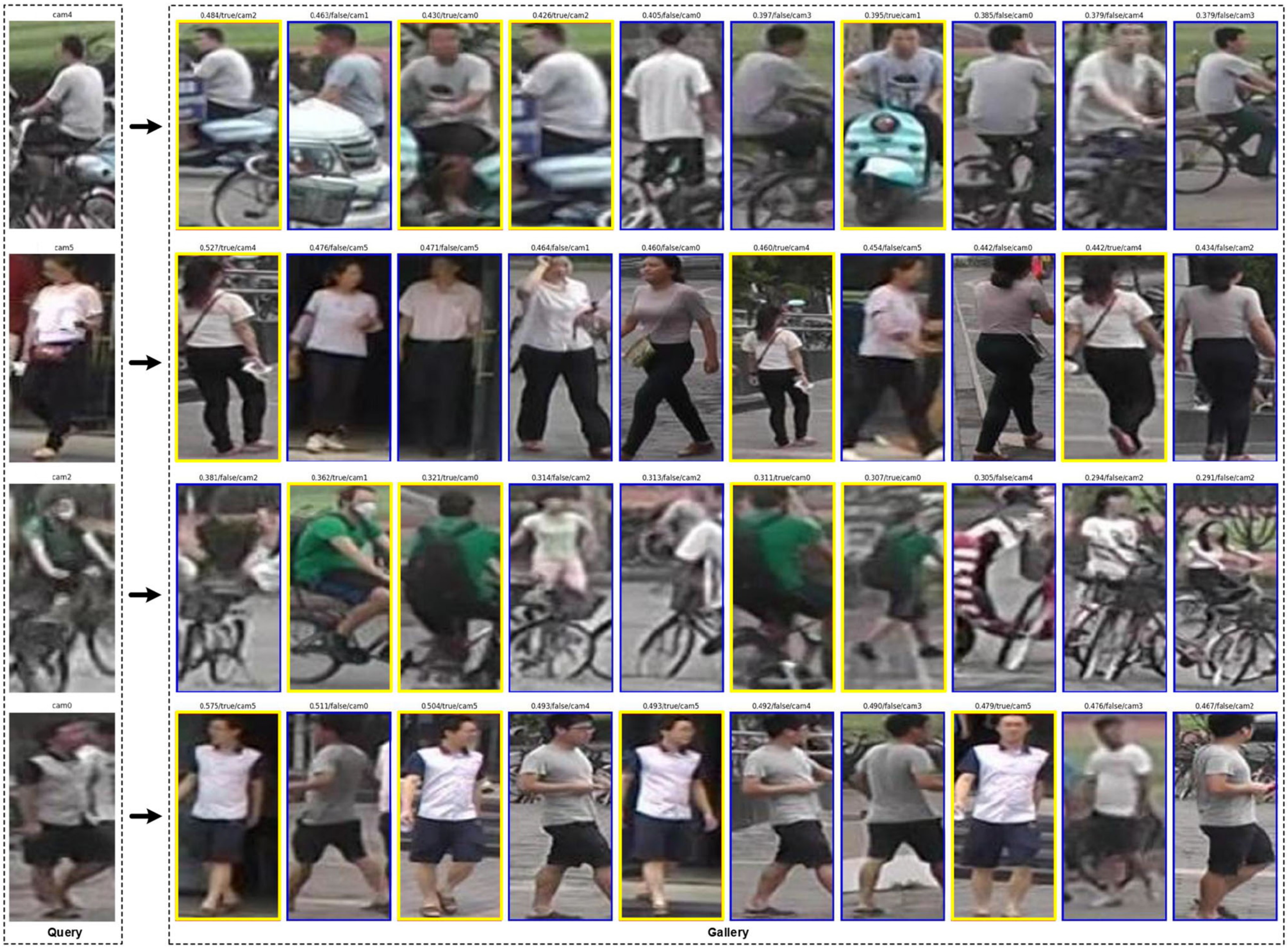

Fig. 9 Some demonstrations of matching results

First of all, it can be seen from Table 10 that, the accuracy of our DFPN model on the three public datasets (namely Market1501, DukeMTMC-reID and Occluded-Duke) has been significantly improved after pretraining. Especially for the mAP indicator, its accuracy increase by 11.94%, 14.79% and 6.74% respectively. What's more, on the Occluded-Duke dataset, the Rank-1 of pretrained DFPN can increase by 8.37% when compared to that without pretraining. This set of experiments comprehensively prove the importance of pretraining. In addition to the above experiments, we also tested the DFPN after deleting IRFB. As shown in last row of Table 10, the accuracy of our model has decreased significantly, which directly proves that the receptive field is one of the starting points to alleviate the occlusion problem.

Then, on the non-occlusion dataset of Market1501 and DukeMTMC-reID, we select five most advanced algorithms to compare with our DFPN (Pretrained). As shown in Table 11, our DFPN is very competitive in terms of the mAP indicator, surpassing CircleLoss, FastReID, CBN and the recently published DeepMiner algorithm (0.96% and 2.57% higher, respectively). For the Rank-1 indicator, the performance of DFPN is very close to that of DeepMiner on the Market1501 dataset, but is about 4% lower than that of DukeMTMC-reID. It is undeniable that when compared with the state-of-the-art FlipReID, our DFPN has a certain gap in terms of Rank-1 or mAP, because the network structure of FlipReID is much more complicated than ours, and its backbone has already exceeded 50 layers.

Finally, on the Occluded-Duke dataset, we also select five most advanced algorithms (top 5) to compare with our DFPN (Pretrained). It should be noted that there are few algorithms tested on Occluded-Duke because it is relatively new. As shown in Table 12, the Rank-1 and mAP of DFPN comprehensively surpass the best Ad-Occluded, SFR and DSR algorithms tested on Occluded-Duke in 2018, and even achieve 4.39% higher on mAP than PGFA which was the top 1 algorithm in 2019. However, for the state-of-the-art HOReID algorithm, which uses complex graph convolution, our DFPN still has 3.38% Rank-1 and 2.11% mAP gaps.

To show the performance of our algorithm more intuitively, we randomly selected 4 Query images, found the same pedestrian images under different cameras in their Gallery, and sorted them according to the similarity from high to low (select the top 10 results). It can be seen from Fig. 9 (highlights indicate the correct matches) that our algorithm can get high rankings of matching the correct pedestrians, which shows good robustness under conditions of blur, occlusion, and different scales.

Conclusion

In this paper, DFPN is proposed to solve the problems of scale, blur and occlusion of pedestrian images in Re-ID tasks. To train faster and more convenient, we use "Shortcut" and the simple design of VGGNet-19 to build backbone RVNet with fewer parameters and no pretraining. After feature extraction, we design a DPFM which combines high-level semantic information and low-level detail information to enhance the richness and representation ability of feature maps. In addition, we introduce the IRFB, which can effectively increase the receptive field of the feature map and reduce parameters. A series of experiments on public datasets have proved the effectiveness of our individual modules and model. In the following work, we will try to further improve our model and apply it in the person search work.

Acknowledgements The authors would like to thank Center for Public Security Information and Equipment Integration Technology, UESTC for providing computation platform.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Zajdel W, Zivkovic Z, Krose BJA (2005) Track of humans: have i seen this person before? In: IEEE International Conference on Robotics and Automation, pp 2081–2086. https://doi.org/10.1109/ROBOT.2005.1570420
2. Khamis S, Kuo C, Singh VK, Shet VD (2014) Joint learning for attribute-consistent person re-identification. European Conference on Computer Vision Workshops 8927:134–146. https://doi.org/10.1007/978-3-319-16199-0_10
3. Zhao R, Ouyang W, Wang X (2013) Person re-identification by salience matching. In: Proceedings of the 2013 IEEE international conference on computer vision (IEEE Computer Society, USA), ICCV '13, pp 2528–2535. https://doi.org/10.1109/ICCV.2013.314
4. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: IEEE International Conference on Computer Vision, pp 3908–3916. https://doi.org/10.1109/CVPR.2015.7299016
5. Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. IEEE conference on computer vision and pattern recognition (CVPR), pp 1249–1258. https://doi.org/10.1109/CVPR.2016.140
6. Chen SZ, Guo CC, Lai JH (2016) Deep ranking for person re-identification via joint representation learning. IEEE Trans Image Process 25(5):2353–2356. https://doi.org/10.1109/TIP.2016.2545929
7. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: deep filter pairing neural network for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 152–159. https://doi.org/10.1109/CVPR.2014.27
8. Wang F, Zuo W, Lin L, Zhang D, Zhang L (2016) Joint learning of single-image and cross-image representations for person re-identification. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1288–1296. https://doi.org/10.1109/CVPR.2016.144
9. Liu H, Feng J, Qi M, Jiang J, Yan S (2017) End-to-end comparative attention networks for person re-identification. IEEE Trans Image Process 26(7):3492–3506. https://doi.org/10.1109/TIP.2017.2700762
10. Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S, Yan J, Wang X (2017) HydraPlus-Net: attentive deep features for pedestrian analysis. In: IEEE International Conference on Computer Vision (ICCV), pp 350–359. https://doi.org/10.1109/ICCV.2017.46
11. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2285–2294. https://doi.org/10.1109/CVPR.2018.00243
12. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person reidentification. arXiv preprint. https://doi.org/10.48550/arXiv.1703.07737
13. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). European Conference on Computer Vision 11208:501–518. https://doi.org/10.1007/978-3-030-01225-0_30
14. Su C, Li J, Zhang S, Xing J, Gao W, Tian Q (2017) Pose-driven deep convolutional model for person re-identification. In: IEEE International Conference on Computer Vision, pp 3960–3969. https://doi.org/10.1109/ICCV.2017.427
15. Wei L, Zhang S, Yao H, Gao W, Tian Q (2019) Glad: global-local-alignment descriptor for scalable person re-identification. IEEE Trans Multimed 21(4):986–999. https://doi.org/10.1109/TMM.2018.2870522
16. Krizhevsky A, Sutskever I, Hinton GE (2017) (2012) Imagenet classification with deep convolutional neural networks. Ann Conf Neural Info Process Syst 2:1097–1105. https://doi.org/10.1145/3065386
17. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint. https://doi.org/10.48550/arXiv.1409.1556
18. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and

Pattern Recognition, pp 770–778. https://doi.org/10.1109/CVPR.2016.90

19. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2261–2269. https://doi.org/10.1109/CVPR.2017.243
20. Srivastava R, Greff K, Schmidhuber J (2017) Highway networks. arXiv preprint. https://doi.org/10.48550/arXiv.1505.00387
21. Huang G, Sun Y, Liu Z, Sedra D, Weinberger KQ (2016) Deep networks with stochastic depth. European Conference on Computer Vision 9908:646–661. https://doi.org/10.1007/978-3-319-46493-0_39
22. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011. https://doi.org/10.1109/TPAMI.2019.2913372
23. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. arXiv preprint. https://doi.org/10.48550/arXiv.1711.07971
24. Cao Y, Xu J, Lin S, Wei F, Hu H (2019) GCNet: non-local networks meet squeeze-excitation networks and beyond. In: IEEE International Conference on Computer Vision Workshop (ICCVW), pp 1971–1980. https://doi.org/10.1109/ICCVW.2019.00246
25. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 936–944. https://doi.org/10.1109/CVPR.2017.106
26. Liu S, Huang D, Wang Y (2018) Receptive field block net for accurate and fast object detection. In: European Conference on Computer Vision. Springer, Cham, vol 11215, pp 404–419. https://doi.org/10.1007/978-3-030-01252-6_24
27. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 1–9. https://doi.org/10.1109/CVPR.2015.7298594
28. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: IEEE Conference on Computer Vision (ICCV), pp 1116–1124. https://doi.org/10.1109/ICCV.2015.133
29. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: Hua G, Jégou H (eds) In: European Conference on Computer Vision. springer, cham, vol 9914, pp 17–35. https://doi.org/10.1007/978-3-319-48881-3_2
30. He L, Liao X, Liu W, Liu X, Cheng P, Mei T (2020) Fastreid: a pytorch toolbox for real-world person re-identification. arXiv preprint. https://doi.org/10.48550/arXiv.2006.02631
31. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 815–823. https://doi.org/10.1109/CVPR.2015.7298682
32. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: IEEE International Conference on Computer Vision (ICCV), pp 1026–1034. https://doi.org/10.1109/ICCV.2015.123
33. Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in convolutional network. arXiv preprint. https://doi.org/10.48550/arXiv.1505.00853
34. Zheng Z, Zheng L, Yang Y (2019) Pedestrian alignment network for large-scale person re-identification. IEEE Trans Circuits Syst Video Technol 29(10):3037–3045. https://doi.org/10.1109/TCSVT.2018.2873599
35. Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. In: IEEE International Conference on Computer Vision (ICCV), pp 3820–3828. https://doi.org/10.1109/ICCV.2017.410
36. Lin Y, Zheng L, Zheng Z, Wu Y, Hu Z, Yan C, Yang Y (2019) Improving person re-identification by attribute and identity learning. Pattern Recognit 95:151–161. https://doi.org/10.1016/j.patcog.2019.06.006
37. Zhong Z, Zheng L, Kang G, Li S, Yang Y (2017) Random erasing data augmentation. arXiv preprint. https://doi.org/10.48550/arXiv.1708.04896
38. Chen Y, Zhu X, Gong S (2017) Person re-identification by deep learning multi-scale representations. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp 2590–2600. https://doi.org/10.1109/ICCVW.2017.304
39. Si J, Zhang H, Li C.G., Kuen J, Kong X, Kot AC, Wang G (2018) Dual attention matching network for context-aware feature sequence based person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 5363–5372. https://doi.org/10.1109/CVPR.2018.00562
40. Tay CP, Roy S, Yap KH (2019) Aanet: attribute attention network for person re-identifications. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 7127–7136. https://doi.org/10.1109/CVPR.2019.00730
41. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in contex. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) European conference on computer vision. Springer, Cham, vol 8693, pp 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
42. Miao J, Wu Y, Liu P, Ding Y, Yang Y (2019) Pose-guided feature alignment for occluded person re-identification. In: International Conference on Computer Vision (ICCV), pp 542–551. https://doi.org/10.1109/ICCV.2019.00063
43. Ni X, Rahtu E (2021) Flipreid: closing the gap between training and inference in person re-identification. arXiv preprint. https://doi.org/10.48550/arXiv.2105.05639
44. Benzine A, Seddik M, Desmarais J (2021) Deep miner: a deep and multi-branch network which mines rich and diverse features for person re-identification. arXiv preprint. https://doi.org/10.48550/arXiv.2102.09321
45. Sun Y, Cheng C, Zhang Y, Zhang C, Zheng L, Wang Z, Wei Y (2020) Circle loss: a unified perspective of pair similarity optimization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 6397–6406. https://doi.org/10.1109/CVPR42600.2020.00643
46. Zhuang Z, Wei L, Xie L, Zhang T, Zhang H, Wu H, Ai H, Tian Q (2020) Rethinking the distribution gap of person re-identification. In: Vedaldi A, Bischof H, Brox T, Frahm JM (eds) European Conference on Computer Vision. Springer, Cham, Vol 12357, pp 140–157. https://doi.org/10.1007/978-3-030-58610-2_9
47. Wang G, Yang S, Liu H, Wang Z, Yang Y, Wang S, Yu G, Zhou E, Sun J (2020) High-order information matters: Learning relation and topology for occluded person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 6448–6457. https://doi.org/10.1109/CVPR42600.2020.00648
48. Huang H, Li D, Zhang Z, et al (2018) Adversarially occluded samples for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 5098–5107. https://doi.org/10.1109/CVPR.2018.00535
49. He L, Liang J, Li H, Sun Z (2018) Recognizing partial biometric patterns. arXiv preprint. https://doi.org/10.48550/arXiv.1810.07399
50. He L, Liang J, Li H, et al (2018) Deep spatial feature reconstruction for partial person re-identification: alignment-free approach. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 7073–7082. https://doi.org/10.1109/CVPR.2018.00739