Integrating Coarse Granularity Part–Level Features with Supervised Global-Level Features for Person Re-Identification

Abstract: Person re-identification (Re-ID) has achieved great progress in recent years. However, person Re-ID methods are still suffering from body part missing and occlusion problems, which makes the learned representations less reliable. In this paper, we propose a robust coarse granularity part-level network (CGPN) for person Re-ID, which extracts robust regional features and integrates supervised global features for pedestrian images. CGPN gains two-fold benefit toward higher accuracy for person Re-ID. On one hand, CGPN learns to extract effective regional features for pedestrian images. On the other hand, compared with extracting global features directly by backbone network, CGPN learns to extract more accurate global features with a supervision strategy. The single model trained on three Re-ID datasets achieves state-of-the-art performances. Especially on CUHK03, the most challenging Re-ID dataset, we obtain a top result of Rank-1/mean average precision (mAP)=87.1%/83.6% without re-ranking.

Keywords: person Re-ID; supervision; coarse granularity

CAO Jiahao[1,2], MAO Xiaofei[1,2], LI Dongfang[1,2], ZHENG Qingfang[1,2], JIA Xia[1,2]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518057, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202101009

http: //kns.cnki.net/kcms/detail/34.1294.TN.20210225.1146.002.html, published online February 25, 2021

Manuscript received: 2020–12–05

Citation (IEEE Format): J.H. Cao, X. F. Mao, D. F. Li, et al., "Integrating coarse granularity part-level features with supervised global-level features for person re-identification," *ZTE Communications*, vol. 19, no. 1, pp. 72 – 81, Mar. 2021. doi: 10.12142/ZTECOM.202101009.

1 Introduction

Person re-identification (Re-ID) aims to retrieve a given person among all the gallery pedestrian images captured by different cameras. It is a challenging task to learn robust pedestrian feature representations as realistic scenarios are highly complicated with regards to the illumination, the background, and occlusion problems. In recent years, person Re-ID has achieved great progress[1 – 7]. However, person Re-ID methods are still suffering from occluded or body part missing pedestrian images, where they fail to extract discriminative deep features for person Re-ID. Intuitively, the complexity of realistic scenarios increases the difficulty in making correct retrieval for person Re-ID[8 – 10]. Therefore, existing person Re-ID methods usually decline a lot in performance when dealing with realistic person Re-ID dataset like CUHK03, which contains a lot of occluded or body part missing pedestrian images, as illustrated in **Fig. 1**.

▲**Figure 1. Pedestrian images in CUHK03-labeled dataset**

As it is well known, part-based methods[5 - 7] like multiple granularity network (MGN)[5] are widely used in person Re-ID and have achieved promising performance. Generally, part-based methods learn to combine global features and discriminative regional features for person Re-ID. The global features in part-based methods are usually extracted directly from the whole person image by the backbone network, while the regional features are generated by directly partitioning feature maps of the whole body into a fixed number of parts. Nevertheless, the overall performance of such part-based methods seriously depends on that all person images are well-bounded holistic person images with few occlusion or body part missing. As real-world scenarios are complicated, the bounding boxes detected by the detection algorithm may not be accurate enough, which usually leads to occluded or body part missing pedestrian images as Fig.1 shows. In Fig.1, ID-A means the ID of the pedestrian is A. We can see that for the same person in the realistic scenario, occluded and body-part missing pedestrian images are both captured as people are moving around the cameras. When dealing with such occluded or body part missing pedestrian images, the global features extracted from the whole image directly by the backbone network become less accurate; moreover, the regional features generated by directly partitioning feature maps of the whole body may focus on occluded parts and become ineffective, which impair the person Re-ID accuracy evidently.

To address the above problems, in this paper, we propose the coarse granularity part-level network (CGPN) for person Re-ID model that learns discriminative and diverse feature representations without using any third models. Our CGPN model can be trained end-to-end and performs well on three person Re-ID datasets. Especially on CUHK03, which contains a lot of occluded or body part missing pedestrian images, our method achieves state-of-the-art performances and outperforms the current best method by a large margin. CGPN has three branches, and each branch consists of a global part and a local part. The global part is supervised to learn more accurate global features by part-level body regions. With the supervision strategy, the global part can learn more proper global features for occluded or body part missing pedestrian images. For the local part, as pedestrian images detected in realistic scenarios are often occluded or body-part missing, too many fine grained local features generated by partitioning the whole body feature maps may decrease model performance. Therefore we propose a coarse grained part-level feature strategy that can extract effective regional features and perform better on the three person Re-ID datasets.

CGPN gains two-fold benefit toward higher accuracy for person Re-ID. Firstly, compared with extracting global features directly by backbone network, CGPN learns to extract more accurate global features with the supervision strategy. Secondly, with the coarse grained part-level feature strategy, CGPN is capable of extracting effective body part features as regional features for person Re-ID. Besides, our method is completely an end-to-end learning process, which is easy for learning and implementation. Experimental results confirm that our method achieves state-of-the-art performances on several mainstream Re-ID datasets, especially on CUHK03, the most challenging dataset for person Re-ID, in single query mode, and we obtain a top result of Rank-1/mean average precision (mAP)=87.1%/83.6% without re-ranking.

The main contributions of our work are summarized as follows:

• We propose a novel framework named CGPN, which effectively integrates coarse grained part-level features and supervised global-level features and is more robust for person Re-ID.

• We develop the coarse grained part-level feature strategy for person Re-ID.

• We prove that the integration model of coarse grained part-level features and supervised global-level features achieves state-of-the-art results on three Re-ID datasets, especially on the CUHK03 dataset, in which our model outperforms the current best method by a large margin.

2 Related Works

2.1 Part-Based Re-ID Model

As deep learning is widely used in person Re-ID nowadays, most existing methods[11 - 12] choose to extract feature maps by directly applying a deep convolution network such as ResNet[13]. However, the single global feature extracted from the whole person image by a deep convolution network does not perform as well as expected. The reason is that person images captured by cameras usually contain random background information and are often occluded or body part missing, which impairs the performance a lot. Then part-based methods are proposed to get additional useful local information from person images for person Re-ID. As an effective way to extract local features, part-based methods[5 - 7, 14] usually benefit from person structure and together with global features, push the performance of person Re-ID to a new level. The common solution of part-based methods is to split the feature maps horizontally into several parts according to human body structure and concatenate the feature maps of each part. However, when dealing with occluded or body part missing pedestrian images, we find that part-based methods like MGN[5], which has received state-of-the-art results on person Re-ID datasets, face the problem of performance decrease. Obviously, part-based methods are common solutions to holistic pedestrian images as they can get correct body parts by uniform partitioning, however, these methods are less effective to occluded or body part missing pedestrian images.

2.2 Attention-Based Re-ID Model

Recently, some attention-based methods try to address the

occlusion or body-part missing problems with the help of attention mechanisms. Attention module is developed to help extract more accurate features by locating the significant body parts and learning the discriminative features from these informative regions. LI et al.[15] propose a part-aligning convolutional neural network (CNN) network for locating latent regions (hard attention) and then extract these regional features for Re-ID. ZHAO et al.[16] employ the spatial transformer network[17] as the hard attention model to find discriminative image parts. LI et al.[18] use multiple spatial attention modules (by softmax function) to extract features at different spatial locations. XU et al.[19] propose to mask the convolutional maps via a pose-guided attention module. LI et al.[14] jointly learn multi-granularity attention selection and feature representation for optimizing person Re-ID in deep learning. However, most of the attention-based methods are often more prone to higher feature correlations, as these methods tend to have features focusing on a more compact subspace, which makes the extracted features attentive but less diverse, and therefore leads to sub-optimal matching performance.

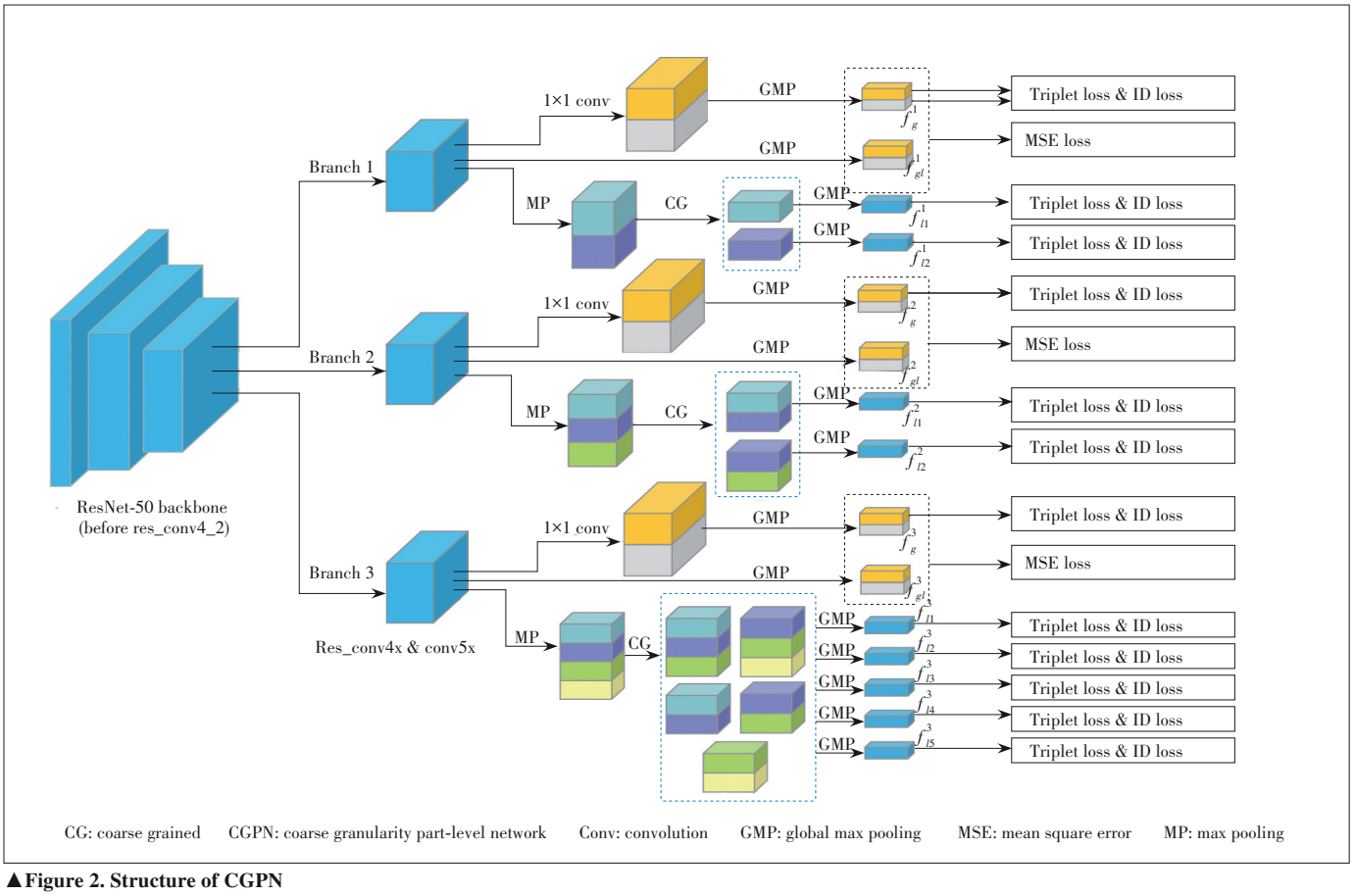
2.3 Pose-Driven Re-ID Model

Some pose-driven methods utilize pose information to tackle the occlusion or body-part missing problems. In these methods, pose landmarks are introduced to help to align body parts as pose landmarks indicate the body position of persons. ZHENG et al.[10] propose to use a CNN-based external pose estimator to normalize person images based on their pose, and the original and normalized images are then used to train a single deep Re-ID embedding. SARFRAZ et al.[20] directly concatenate fourteen landmarks confidence maps with the image as network input, letting the model automatically learn alignment way. HUANG et al.[21] propose a part aligned pooling (PAP) that utilizes seventeen human landmarks to enhance alignment. MIAO et al.[22] learn to exploit pose landmarks to disentangle the useful information from the occlusion noise. However, the landmarks of persons are obtained usually by a third pose estimation model trained on an extra dataset, which increases the complicity of the whole Re-ID network. What's more, standard pose estimation datasets may not cover the drastic viewpoint variations in surveillance scenarios, and besides, surveillance images may not have sufficient resolution for stable landmarks prediction.

3 Proposed Method

3.1 Structure of CGPN

In this part, we present our CGPN structure in **Fig. 2**, in



▲**Figure 2. Structure of CGPN**

which the ResNet-50 backbone is split into three branches after res_conv4_1 block. Each branch consists of a global part and a local part. In the global part, we apply two 1×1 convolutional layers and global max pooling (GMP) to generate global features. While in the local part, we apply a max pooling (MP) with different kernel sizes, split the feature maps into different spatial horizontal stripes, and then apply a coarse grained (CG) strategy and GMP to generate local features. The backbone of our network is a CNN structure, such as ResNet[13], which achieves competitive results in many deep learning tasks. Like MGN[5], we divide the output of res_conv4_1 into three different branches. Through the backbone network, CGPN transfers the input image into a 3D tensor $\boldsymbol{T}$ with size of $c\times h\times w$ (c is the channel number, h is the height, and w is the width). Each of the three branches contains a global part and a local part. The global part in three branches shares the same structure, while in every local part the output feature maps are uniformly partitioned into different stripes.

In the global part, two 1×1 convolution layers are applied to the output feature maps to extract regional features. Each of the 1×1 convolution layers will output c-channel features and be supervised by the corresponding part features. In futher detail, for i-th branch's global part, to supervise the global features, the output feature maps are uniformly divided into two parts in the vertical direction, and a global pooling is applied to each of them to extract two part features $\{f_{gl1}^{\ i}, f_{gl2}^{\ i}\}$. The two part features $\{f_{gl1}^{\ i}, f_{gl2}^{\ i}\}$ are utilized in the training stage to supervise global features $\{f_{g1}^{\ i}, f_{g2}^{\ i}\}$ generated by the two 1×1 convolution layers. After the training stage finishes, the two part features are no longer needed. The first c-channel global features $f_{g1}^{\ i}$ should be closer to the upper part features $f_{gl1}^{\ i}$, and in the same way, the second c-channel global features $f_{g2}^{\ i}$ should be closer to the bottom part features $f_{gl2}^{\ i}$. In the inference stage, the first c-channel global features $f_{g1}^{\ i}$ and the second c-channel global features $f_{g2}^{\ i}$ are concatenated to form 2 c-channel features as final global features $f_g^{\ i}$. As the global parts of the three branches all share the same structure, we can get three global features $\{f_g^{\ 1}, f_g^{\ 2}, f_g^{\ 3}\}$ in total. With the supervision of the part features, in the final 2 c-channel global features, the first c-channel global features are forced to focus on the upper part of the human body, while the second c-channel global features focus on the bottom part of the human body, which makes final global features more robust to person image occlusion or body-part missing.

For the local part in three branches, the output feature maps are divided into N stripes in the vertical direction with each stripe having the size of $c\times(h/N)\times w$, from which we prepare to extract local features. However, for person images that are occluded or body part missing, it might be harmful and decrease the performance of person Re-ID, if the granularity of local features is too fine. To alleviate the drawbacks of fine-grained local features, we choose to extract local features in a bigger receptive field that contains enough body structure information to well represent the corresponding body region. In this paper, we propose a coarse-grained part-level feature strategy in which the combined part stripes must be adjacent and the minimum height proportion of combined local features should be no less than half of the output feature maps. The detail of the coarse grained strategy is illustrated in **Fig. 3.** In the local part of the first branch, the output feature maps are divided into two stripes in vertical direction as shown in Fig. 3a, and then pooling operations are performed to get local feature representations $\{f_{l1}^{\ 1}, f_{l2}^{\ 1}\}$ corresponding to the size of $c\times(h/2)\times w$. In the local part of the second branch, the output feature maps are divided into three stripes but we combine two adjacent stripes to get two 2/3 proportion local features $\{f_{l1}^{\ 2}, f_{l2}^{\ 2}\}$ corresponding to the size of $c\times(2h/3)\times w$. For the local part of the third branch, the output feature maps are divided into four stripes, and then we combine two and three adjacent stripes to get three 1/2 proportion and two 2/3 proportion local features respectively, with $\{f_{l1}^{\ 3}, f_{l2}^{\ 3}, f_{l3}^{\ 3}, f_{l4}^{\ 3}, f_{l5}^{\ 3}\}$ corresponding to the size of $c\times(3h/4)\times w$, $c\times(3h/4)\times w$, $c\times(3h/4)\times w$, $c\times(h/2)\times w$, $c\times(h/2)\times w$ respectively.

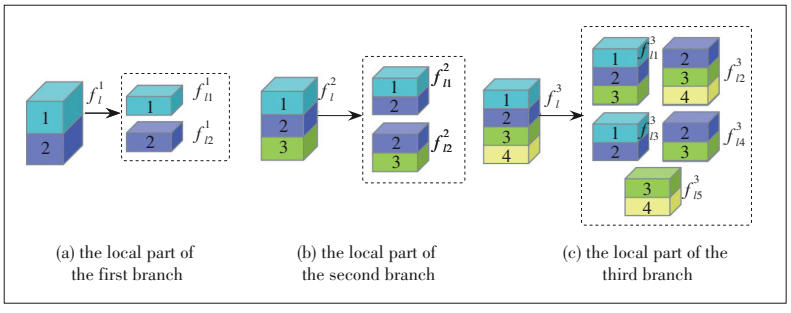
During the test, all features $\{f_g^{\ 1}, f_g^{\ 2}, f_g^{\ 3}, f_{l1}^{\ 1}, f_{l2}^{\ 1}, f_{l1}^{\ 2}, f_{l2}^{\ 2}, f_{l1}^{\ 3}, f_{l2}^{\ 3}, f_{l3}^{\ 3}, f_{l4}^{\ 3}, f_{l5}^{\ 3}\}$ generated by the global part and the local part in each branch are reduced to 256-dimension and are concatenated together as the final features, as different branches in CGPN actually learn representing information with different granularities which can cooperatively supplement discriminating information after the concatenation operation.

3.2 Loss Functions

Like various deep Re-ID methods, we employ softmax loss for classification, and triplet loss[23] for metric learning. For the supervision of the global part in each branch, we use mean square error (MSE) loss in the training stage.

To be precise, in each branch, the local part is trained with the combination of softmax loss and triplet loss while the global part is trained with MSE loss, softmax loss and triplet loss as illustrated in Fig. 2.

For i-th learned features f_i, $\boldsymbol{W}_k$ is a weight vector for class k with the total class number C. N is the number of training examples in a mini-batch, and the softmax loss is formulated as:



▲**Figure 3. Coarse-grained part-level feature strategy**

$$L_{\text{softmax}} = -\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T f_i}}{\sum_{k=1}^{C} e^{W_k^T f_i}}. \quad (1)$$

We employ the softmax loss to all global features $\{f_g^1, f_g^2, f_g^3\}$, and all coarse grained local features $\left\{f_{li}^1 \middle|_{i=1}^{2}, f_{li}^2 \middle|_{i=1}^{2}, f_{li}^3 \middle|_{i=1}^{5}\right\}$.

Besides, all the global features $\{f_g^1, f_g^2, f_g^3\}$ are also trained with triplet loss. In the training stage, an improved batch hard triplet loss is applied with formula as follows:

$$L_{\text{triplet}} = \sum_{i=1}^{P}\sum_{a=1}^{K}\left[\alpha + \max_{p=1,\ldots,K}\left\|f_a^{(i)} - f_p^{(i)}\right\|_2 - \min_{\substack{n=1,\ldots,K \\ j=1,\ldots,P \\ j\neq i}}\left\|f_a^{(i)} - f_n^{(j)}\right\|_2\right]_+ . \quad (2)$$

In the above formula, P is the number of selected identities and K is the number of images from each identity in a mini-batch. f_a^i is the anchor sample, f_p^i is the positive sample, f_n^i is the negative sample and α is the margin parameter to control the differences of intra- and inter-distances, which is set to 1.2 in our implementation.

To supervise the global features, we employ the MSE loss to all global features $\{f_g^1, f_g^2, f_g^3\}$ and the supervision local features $\{f_{gl}^1, f_{gl}^2, f_{gl}^3\}$ with the formula as follows:

$$L_{\text{mse}} = \sum_{i=1}^{B}\sum_{p=1}^{M}\left\|f_{gp}^i - f_{glp}^i\right\|_2^2, \quad (3)$$

where f_{gp}^i is the p-th c-channel features of global features in i-th branch, f_{glp}^i is the supervision p-th c-channel part features in the same branch. As the global part consists of two c-channel features and there are three branches in our network, M is set to 2 and B is set to 3 in our implementation.

The overall training loss is the sum of above three losses, which is formulated by:

$$L = L_{\text{softmax}} + L_{\text{triplet}} + L_{\text{mse}}. \quad (4)$$

4 Experiment

4.1 Datasets and Protocols

We train and test our model respectively on 4 mainstream Re-ID datasets: Market-1501[24], DukeMTMC-reID[25], CUHK03[26] and Occluded-DukeMTMC[22]. Especially the CUHK03 dataset, which is the most challenging realistic scenario Re-ID dataset as it consists of a lot of occluded or body part missing pedestrian images as illustrated in Fig. 1.

Market-1501 is captured by six cameras in front of a campus supermarket, which contains 1 501 person identities, 12 936 training images from 751 identities and 19 732 testing images from 750 identities. All provided pedestrian bounding boxes are detected by deformable part models (DPM)[27].

DukeMTMC-reID contains 1 812 person identities captured by 8 high-resolution cameras. There are 1 404 identities in more than two cameras and the other 408 identities are regarded as distractors. The training set consists of 16 552 images from 702 identities and the testing set contains 17 661 images from the rest 702 identities.

CUHK03 contains 1 467 person identities captured by six cameras on campus of The Chinese University of Hong Kong (CUHK). Both manually labeled pedestrian bounding boxes and automatically detected bounding boxes are provided. In this paper, we use the manually labeled version and follow the new training/testing protocol proposed in Ref. [28], with 7 368 images from 767 identities for training and 5 328 images from 700 identities for testing.

Occluded-DukeMTMC is re-segmented from the original DukeMTMC-reID dataset. The training set contains 15 618 images, and the gallery set and query set contain 17 661 and 2 210 images, respectively, in which all query images and some gallery images are occluded images, and these occluded images retain their occluded regions without being manually cropped.

In our experiment, we report the average cumulative match characteristic (CMC) at Rank-1, Rank-5, Rank-10 and mean average precision (mAP) on all the candidate datasets to evaluate our method.

4.2 Implementation Details

All images are re-sized into 384×128 px and the backbone network is ResNet-50[13], pre-trained on ImageNet with the original fully connected layer discarded. In the training stage, the mini-batch size is set to 64, in which we randomly select 8 identities and 8 images for each identity (P=8, K=8). Besides, we deploy a randomly horizontal flipping strategy to images for data augmentation. Different branches in the network are all initialized with the same pre-trained weights of corresponding layers after res_conv4_1 block. Our model is implemented on Pytorch platform. We use stochastic gradient descent (SGD) as the optimizer with the default hyper-parameters (momentum= 0.9, weight decay factor=0.0005) to minimize the network loss. The initial learning rate is set to 1e-2 and we decay it at epoch 60 and 80 to 1e-3 and 1e-4 respectively. The total training takes 240 epochs. During the evaluation, we use the average of original image features and horizontally flipped image features as the final features. All of our experiments on different datasets follow the settings above.

4.3 Comparison with State-of-the-Art Methods

In this section, we compare our proposed approach with current state-of-the-art methods on the three main-stream Re-ID datasets.

The statistical comparison between our PGCN network and

the state-of-the-art methods on Market-1501, DukeMTMC-reID and CUHK03 datasets is shown in **Table 1**.

On Market-1501 dataset, semantics aligning network (SAN) achieves the best published result without re-ranking, but our CGPN achieves 89.9% on the metric mAP, exceeding SAN by +1.9%. On the metric Rank-1, our CGPN achieves 96.1%, on a par with SAN, while our model is trained in an easier and end-to-end way. Compared with multiple granularity network (MGN) which is also a multiple branches method, our model surpasses MGN by +0.4% on the metric Rank-1 and by +3.0% on the metric mAP.

Among the performance comparisons on DukeMTMC-reID dataset, Pyramid achieved the best published result on metrics Rank-1 and mAP respectively. Our CGPN achieves the state-of-the-art result of Rank-1/mAP = 90.4%/80.9%, outperforming Pyramid by +1.4% on the metric Rank-1 and +1.9% on the metric mAP.

From Table 1, our CGPN model achieves Rank-1/mAP = 87.1%/83.6% on the most challenging CUHK03 labeled dataset under the new protocol. On the metric Rank-1, our CGPN outperforms the best published result of SAN by +7.0% and outperforms the best published result of Pyramid by +6.7% on mAP.

▼Table 1. Performance comparisons with the state-of-the-art results on Market-1501, DukeMTMC-reID and CUHK03 datasets in single query mode without re-ranking

Method	Market-1501		DukeMTMC-reID		CUHK03	
	Rank-1/%	mAP/%	Rank-1/%	mAP/%	Rank-1/%	mAP/%
IDE[29]	-	-	-	-	22.2	21.0
PAN[30]	-	-	-	-	36.9	35.0
SVDNet[31]	-	-	-	-	40.9	37.8
IDE(R)+DM[3][32]	73.4	51.8	-	-	-	-
MGCAM[33]	83.8	74.3	-	-	50.1	50.2
DHA-Net + ISO(Aggr)[34]	88.2	70.1	74.2	54.5	-	-
HA-CNN[14]	91.2	75.7	80.5	63.8	44.4	41.0
VPM[35]	93.0	80.8	83.6	72.6	-	-
SCP[36]	94.1	-	84.8	-	-	-
PCB+RPP[6]	93.8	81.6	83.3	69.2	-	-
SphereReID[37]	94.4	83.6	83.9	68.5	-	-
MGN[5]	95.7	86.9	88.7	78.4	68.0	67.4
DSA[38]	95.7	87.6	86.2	74.3	78.9	75.2
Pyramid[7]	95.7	88.2	89.0	79.0	78.9	76.9
SAN[39]	96.1	88.0	87.9	75.5	80.1	76.4
CGPN	96.1	89.9	90.4	80.9	87.1	83.6

CGPN: coarse granularity part-level network
DHA: deep hidden attribute
DM: discrepancy matrix and matrix metric
DSA: densely Semantically Aligned
HA-CNN: harmonious attention convolutional neural network
IDE: ID-discriminative embedding
ISO: Identity-preserving, Sparsity constraints and the Orthogonal generation module
mAP: mean average precision
MGCAM: mask-guided contrastive attention model
MGN: multiple granularity network
PAN: pedestrian alignment network
PCB: part-based convolutional baseline
RPP: refined part pooling
SAN: semantics aligning network
SCP: spatial-channel parallelism
SVDNet: singular vector decomposition network
VPM: visibility-aware part model

In summary, our proposed CGPN can always outperform all other existing methods and shows strong robustness over different Re-ID datasets. According to the comparative experiments on the three datasets, especially on CUHK03 dataset, our approach can consistently outperform all other models by a large margin. Therefore, we can conclude that our method can effectively extract robust deep features from occluded or body part missing pedestrian images in person Re-ID.

We also conduct an experiment and compare the performances with the existing methods on Occluded-DukeMTMC. The results are listed in **Table 2**. As can be seen that, CGPN gets the top performance among the compared approaches, and obtains 58.5%/50.9% in rank-1/mAP. CGPN surpasses pose-guided feature alignment (PGFA) by +7.1% rank-1 accuracy and +13.6% mAP, which is a large margin. Therefore, we can conclude that our proposed CGPN integrated with supervised global-level features can effectively address the occlusion problem in person Re-ID.

4.4 Importance of Coarse-Grained Part-Level Features

To verify the effectiveness of coarse-grained part-level feature strategy in the CGPN model, we train two mal-functioned CGPN for comparison:

• CGPN-1 abandons the local parts in three branches and keeps only global parts.

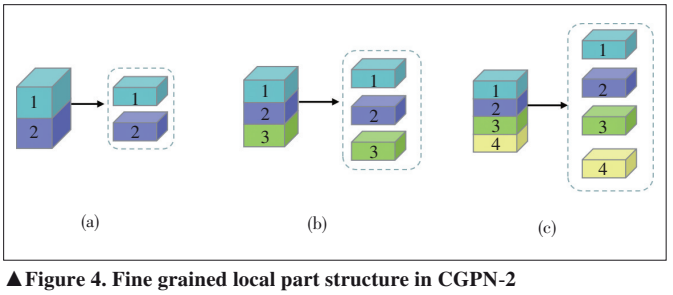
• CGPN-2 replaces coarse-grained part-level features with fine-grained part-level features. It abandons coarse-grained strategy in local parts of three branches, compared with the normal CGPN model. Its local parts in three branches directly divide the output feature maps into two, three and four parts as shown in **Fig. 4**.

From the comparison of CGPN-1 with CGPN, we can see

▼Table 2. Performance comparisons with the state-of-the-art results on Occluded-DukeMTMC dataset in single query mode without re-ranking.

Method	Occluded-DukeMTMC			
	Rank-1/%	Rank-5/%	Rank-10/%	mAP/%
HA-CNN[14]	34.4	51.9	59.4	26.0
PCB+RPP[6]	42.6	57.1	62.9	33.7
PGFA[22]	51.4	68.6	74.9	37.3
CGPN	58.5	73.4	78.4	50.9

CGPN: coarse granularity part-level person Re-ID network
HA-CNN: harmonious attention convolutional neural network
mAP: mean average precision
PCB: part-based convolutional baseline
PGFA: pose-guided feature alignment
RPP: refined part pooling



▲Figure 4. Fine grained local part structure in CGPN-2

a significant performance decrease on Rank-1/mAP by −1.2%/−2.0%, −1.1%/−2.5% and −4.7%/−3.7% on Market-1501, DukeMTMC-reID and CUHK03 datasets respectively. Especially on CUHK03, we can observe a sharp decrease by −4.7%/−3.7% on the metric Rank-1/mAP. As CGPN-1 is trained in exactly the same procedure with the CGPN model and the CUHK03 dataset typically consists of many occluded or body part missing person images, we can infer that the coarse-grained local part is critical for CGPN model, especially on the dataset which contains a lot of occluded or body part missing person images.

Comparing CGPN-2 with CGPN, we can still observe a performance decrease by −0.8%/−0.5%, −0.1%/−0.7% and −1.8%/−1.1% on the metric Rank-1/mAP on Market-1501, DukeMTMC-reID and CUHK03 datasets respectively. Compared with fine-grained part-level features, coarse-grained part-level features contain enough body structure information to better represent the corresponding body regions, which makes CGPN learn more robust local features. Besides, on CUHK03, we can also see a sharper performance decrease compared with the other two datasets. The reason is that Market-1501 and DukeMTMC-reID consist of mainly holistic person images with little occlusions or body part missing, and these images keep complete body structure and make fine-grained part-level features achieve comparable performance with coarse-grained part-level features. While on CUHK03, as it consists of a lot of occluded or body part missing person images, coarse-grained part-level features outperform fine-grained part-level features evidently. Our experiments clearly prove that our coarse- grained part-level feature strategy can improve model performance significantly and is critical for model robustness, especially for occluded or body part missing person images.

4.5 Importance of Supervised Global Part

To further verify the effectiveness of the supervised global part in CGPN model, we train another two mal-functioned CGPN for comparison:

• CGPN-3 abandons global parts in all three branches and keeps only local parts that are trained with triplet loss and softmax loss.

▼ Table 3. Ablation study of CGPN coarse grained part-level feature strategy and supervised global part, with comparison results on Market-1501, DukeMTMC-reID and CUHK03-labeled at evaluation metrics of Rank-1 and mAP in single query mode without re-ranking

Method	Market-1501		DuckMTMC-reID		CUHK03	
	Rank-1/%	mAP/%	Rank-1/%	mAP/%	Rank-1/%	mAP/%
CGPN-1	94.9	87.9	89.3	78.4	82.4	79.9
CGPN-2	95.3	89.4	90.3	80.2	85.3	82.5
CGPN-3	94.2	86.2	88.6	76.9	84.3	81.0
CGPN-4	95.2	89.3	90.0	79.9	83.4	80.7
CGPN	96.1	89.9	90.4	80.9	87.1	83.6

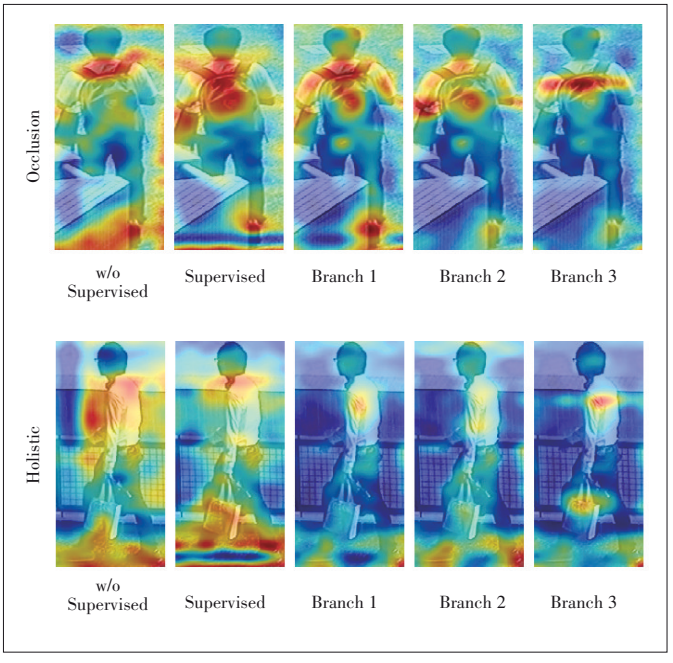
CGPN: coarse granularity part-level network mAP: mean average precision

• CGPN-4 keeps the global parts but abandons the supervision learning of all global parts in three branches, and these global parts are trained only with triplet loss and softmax loss.

Comparing CGPN-3 with CGPN, we observe a dramatic performance decrease on all three datasets. The performance on the metric Rank-1/mAP decreases by −1.9%/−3.7%, −1.8%/−4.0% and −2.8%/−2.6% on Market-1501, DukeMTMC-reID and CUHK03 respectively. As the three models are trained in exactly the same procedure, we conclude that the global part is critical to CGPN.

Comparing CGPN-4 with CGPN, after abandoning global supervision, we observe a performance decrease on Rank-1/mAP of −0.9%/−0.6% and −0.4%/−1.0% on Market-1501 and DukeMTMC-reID. While on CUHK03 we observe a dramatic performance decrease by −3.7%/−2.9%. The reason of such a different performance decrease is that Market-1501 and DukeMTMC-reID mainly consist of holistic person images from which the global part can get enough good global features directly even without supervision, while CUHK03 contains a lot of occluded or body part missing person images and the supervised global part is much more important for extracting accurate global features.

Comparing CGPN-4 with CGPN-3, after adding unsupervised global parts, we see a large performance improvement on Rank-1/mAP of +1.0%/+3.1% and +1.4%/+3.0% on Market-1501, DukeMTMC-reID. But on CUHK03 we observe a significant performance decrease by −0.9%/−0.3% unexpectedly. As analyzed above, the image type is quite different in the three datasets, especially CUHK03 which contains a lot of occluded or body part missing person images.



▲ Figure 5. Feature visualization of different branches in the two cases of occluded pedestrian images and holistic pedestrian images

The unexpected performance decrease on CUHK03 further proves that unsupervised global features can be harmful and certainly impair model performance. We conclude that the supervision of global features is critical for high performance of person Re-ID and that unsupervised global features will result in inaccurate global features which impair model performance evidently. As shown in **Fig. 5**, we can find that the unsupervised global features may receive interference from the background or occlusion. But, our proposed supervised global features are more robust to person image occlusion or body-part missing.

4.6 Branch Settings Ablation Study

In this section, we conduct a large number of comparative experiments on CUHK03 dataset to verify the effectiveness of the numbers of 1×1 convolution layers in the global part and multi-branch architecture settings.

In the global part, the number of 1×1 convolution layers is a hyper-parameter and influences the receptive field of its corresponding supervision part features. To evaluate the effect of various numbers of 1×1 convolution layers in the global part, as three branches' global part all share the same structure, we only keep branch 1 of CGPN and abandon the other two branches of CGPN. We also abandon the local part of branch 1 and only keep the global part of branch 1 denoted as Branch1-Global.

Table 4 shows the results of Branch1-Global with different numbers of 1×1 convolution layers, i.e. 2, 3, 4, 8. From these results, we can find that Branch1-Global reaches the best performance with two 1×1 convolution layers, which achieves Rank-1/mAP = 77.5%/74.7% on the CUHK03 dataset. The experiment results further illustrate that coarse grained features can make full use of local information and preserve more semantic information, and thus help to extract more accurate global features. Therefore, We finally adopt two 1×1 convolution layers in our CGPN architecture.

We further perform experiments to verify the importance of various branch settings in CGPN. Here, Branch x means we only keep CGPN's backbone and branch x after res_conv4_1. For example, Branch 1 means just preserving CGPN's backbone and branch 1 of CGPN and removing branches 2 and branch 3. With the increasing number of branches, Rank-1/mAP is significantly improved from 82.9%/79.4% to 85.4%/82.2% even to 87.1%/83.6%, as illustrated in Table 4. But, when we try more branches, such as CGPN + Branch 4 and CGPN + Branch 4 + Branch 5, we observe a significant performance decrease on Rank-1/mAP of −1.7%/−1.2% and −1.3%/−1.0% unexpectedly. Therefore, we can conclude that the carefully-designed net-

▼Table 4. Comparison results of different number of 1×1 convolution layers in the global part and multi-branch settings on CUHK03 dataset at evaluation metrics of Rank-1 and mAP in single query mode without re-ranking

Model	Rank-1/%	mAP/%
Branch1-Global w/2-part supervised	77.5	74.7
Branch1-Global w/3-part supervised	78.2	74.5
Branch1-Global w/4-part supervised	76.6	73.4
Branch1-Global w/8-part supervised	76.1	73.5
Branch 1	82.9	79.4
Branch 2	81.8	79.2
Branch 3	82.6	78.7
Branch 2 & Branch 3	84.5	82.1
Branch 1 & Branch 3	83.6	81.1
Branch 1 & Branch 2	85.4	82.2
CGPN + Branch 4	85.4	82.4
CGPN + Branch 4 + Branch 5	85.8	82.6
CGPN	87.1	83.6

CGPN: coarse granularity part-level network mAP: mean average precision

▲Figure 6. Top-4 ranking list for some query images on CUHK03-labeled dataset by CGPN and MGN

work architecture is also the main contributor to performance improvement, and three branches can effectively and efficiently capture enough complement information.

Besides, from Fig. 5, we can find that the three branches of CGPN focus on different parts of the pedestrian, and the extracted features are complementary to each other.

4.7 Visualization of Re-ID Results

We visualize the retrieval results by CGPN and MGN for some given query pedestrian images of CUHK03-labeled dataset in **Fig. 6**, in which the retrieved images are all from the gallery set, but not from the same camera shot. The images with green borders belong to the same identity as the given query, and those with red borders do not. These retrieval results show the great robustness of our CGPN model, regardless of the occlusions or body part missing of detected pedestrian images. CGPN can robustly extract discriminative features for different identities.

5 Conclusions

In this paper, we propose a coarse-grained part-level features learning network integrated with supervised global-level features for person Re-ID. With the coarse-grained part-level strategy, the local parts in three branches learn more discriminative local features. With the supervision learning of global parts in three branches, the global parts learn to extract more accurate and suitable global features for pedestrian images. Experiments have confirmed that our model not only achieves state-of-the-art results on all three mainstream person Re-ID datasets, but pushes the performance to an exceptional level.

References

[1] CHANG X B, HOSPEDALES T M, XIANG T. Multi-level factorisation net for person re-identification [EB/OL]. (2018-04-17) [2020-12-05]. https://arxiv.org/abs/1803.09132
[2] LIU H, FENG J, QI M, et al. End-to-end comparative attention networks for person re-identification [J]. IEEE transactions on image processing, 2017, 26(7): 3492 – 3506. DOI: 10.1109/tip.2017.2700762
[3] SARFRAZ M S, SCHUMANN A, EBERLE A, et al. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 420 – 429. DOI: 10.1109/cvpr.2018.00051
[4] SHEN Y T, LI H S, XIAO T, et al. Deep group-shuffling random walk for person re-identification [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 2265 – 2274. DOI: 10.1109/CVPR.2018.00241
[5] WANG G S, YUAN Y F, CHEN X, et al. Learning discriminative features with multiple granularities for person re-identification [C]//Proceedings of the 26th ACM International Conference On Multimedia. Seoul, Korea: ACM, 2018: 274 – 282. DOI: 10.1145/3240508.3240552
[6] SUN Y F, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [C]//Proceedings of the European Conference on Computer Vision. Munich, German: ECCV, 2018: 480 – 496. DOI: 10.1007/978-3-030-01225-0_30
[7] ZHENG F, DENG C, SUN X, et al. Pyramidal person re-identification via multi-loss dynamic training [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: CVPR, 2019: 8514 – 8522. DOI: 10.1109/cvpr.2019.00871
[8] SHEN Y, LIN W, YAN J, et al. Person re-identification with correspondence structure learning [C]//Proceedings of the IEEE international conference on computer vision. Santiago, Chile: IEEE, 2015: 3200 – 3208. DOI: 10.1109/iccv.2015.366
[9] VARIOR R R, SHUAI B, LU J W, et al. A siamese long short-term memory architecture for human re-identification [C]//European Conference on Computer Vision. Amsterdam, Netherlands, ECCV, 2016: 135 – 153. DOI: 10.1007/978-3-319-46478-7_9
[10] ZHENG L, HUANG Y J, LU H C, et al. Pose-invariant embedding for deep person re-identification [J]. IEEE transactions on image processing, 2019, 28 (9): 4500 – 4509. DOI: 10.1109/tip.2019.2910414
[11] LI W, ZHAO R, XIAO T, et al. DeepReID: deep filter pairing neural network for person re-identification [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014: 152 – 159. DOI: 10.1109/CVPR.2014.27
[12] YI D, LEI Z, LIAO S C, et al. Deep metric learning for person re-identification [C]//2014 22nd International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014: 34 – 39. DOI: 10.1109/icpr.2014.16
[13] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: CVPR, 2016: 770 – 778. DOI: 10.1109/cvpr.2016.90
[14] LI W, ZHU X T, GONG S G. Harmonious attention network for person re-identification [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: CVPR, 2018: 2285 – 2294. DOI: 10.1109/cvpr.2018.00243
[15] LI D W, CHEN X T, ZHANG Z, et al. Learning deep context-aware features over body and latent parts for person re-identification [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: CVPR, 2017: 7398 – 7407. DOI: 10.1109/CVPR.2017.782
[16] ZHAO L M, LI X, ZHUANG Y T, et al. Deeply-learned part-aligned representations for person re-identification [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 3219 – 3228. DOI: 10.1109/iccv.2017.349
[17] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 2017 – 2025
[18] LI S, BAK S, CARR P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification [C]//2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 369 – 378. DOI: 10.1109/CVPR.2018.00046
[19] XU J, ZHAO R, ZHU F, et al. Attention-aware compositional network for person re-identification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: CVPR, 2018: 2119 – 2128. DOI: 10.1109/cvpr.2018.00226
[20] SARFRAZ M S, SCHUMANN A, EBERLE A, et al. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking [C]//2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: CVPR, 2018: 420 – 429. DOI: 10.1109/CVPR.2018.00051
[21] HUANG H J, YANG W J, CHEN X T, et al. EANet: enhancing alignment for cross-domain person re-identification [EB/OL]. (2018-12-19) [2020-12-05]. https://arxiv.org/abs/1812.11369
[22] MIAO J X, WU Y, LIU P, et al. Pose-guided feature alignment for occluded person re-identification [C]//2019 IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 542 – 551. DOI: 10.1109/ICCV.2019.00063
[23] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification [EB/OL]. (2017-03-22) [2020-12-12]. https://arxiv.org/abs/1703.07737v4
[24] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: A benchmark [C]//2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1116 – 1124. DOI: 10.1109/ICCV.2015.133
[25] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set

for multi-target, multi-camera tracking [C]//European Conference on Computer Vision. Amsterdam, Netherlands: ECCV, 2016: 17 – 35. DOI: 10.1007/978-3-319-48881-3_2
[26] LI W, ZHAO R, XIAO T, et al. DeepReID: deep filter pairing neural network for person re-identification [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014: 152 – 159. DOI: 10.1109/CVPR.2014.27
[27] FELZENSZWALB P, MCALLESTER D, RAMANAN D. discriminatively trainedA, multiscale, deformable part model [C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2008: 1 – 8. DOI: 10.1109/CVPR.2008.4587597`
[28] ZHONG Z, ZHENG L, CAO D L, et al. Re-ranking person re-identification with k-reciprocal encoding [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 3652 – 3661. DOI: 10.1109/CVPR.2017.389
[29] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: past, present and future [EB/OL]. [2020-12-05]. https://www.arxiv-vanity.com/papers/1610.02984/
[30] ZHENG Z D, ZHENG L, YANG Y. Pedestrian alignment network for large-scale person re-identification [J]. IEEE transactions on circuits and systems for video technology, 2019, 29(10): 3037 – 3045. DOI: 10.1109/TCSVT.2018.2873599
[31] SUN Y F, ZHENG L, DENG W J, et al. SVDNet for pedestrian retrieval [C]//2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 3820 – 3828. DOI: 10.1109/ICCV.2017.410
[32] WANG Z, HU R M, CHEN C, et al. Person re-identification via discrepancy matrix and matrix metric [J]. IEEE transactions on cybernetics, 2018, 48(10): 3006-3020. DOI: 10.1109/TCYB.2017.2755044
[33] SONG C F, HUANG Y, OUYANG W L, et al. Mask-guided contrastive attention model for person re-identification [C]//2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 1179 – 1188. DOI: 10.1109/CVPR.2018.00129
[34] WANG Z, JIANG J J, WU Y, et al. Learning sparse and identity-preserved hidden attributes for person re-identification [J]. IEEE transactions on image processing, 2019, 29: 2013-2025. DOI: 10.1109/TIP.2019.2946975
[35] SUN Y F, XU Q, LI Y L, et al. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification [C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 393 – 402. DOI: 10.1109/CVPR.2019.00048
[36] FAN X, LUO H, ZHANG X, et al. SCPNet: spatial-channel parallelism network for joint holistic and partial person re-identification [C]//Asian Conference on Computer Vision. Perth, Australia: ACCV, 2018: 19 – 34. DOI: 10.1007/978-3-030-20890-5_2
[37] FAN X, JIANG W, LUO H, et al. SphereReID: Deep hypersphere manifold embedding for person re-identification [J]. Journal of visual communication and image representation, 2019, 60: 51 – 58. DOI: 10.1016/j.jvcir.2019.01.010
[38] ZHANG Z Z, LAN C L, ZENG W J, et al. Densely semantically aligned person re-identification [C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 667 – 676. DOI: 10.1109/CVPR.2019.00076
[39] JIN X, LAN C L, ZENG W J, et al. Semantics-aligned representation learning for person re-identification [EB/OL]. (2019-05-30) [2020-12-05]. https://arxiv.org/abs/1905.13143

Biographies

CAO Jiahao (cao.jiahao@zte.com.cn) received the M.S. degree from Northeastern University, China in 2019 and joined ZTE corporation after he graduated. His current research interests include image processing and deep learning technologies.

MAO Xiaofei received the M.S. degree from TELECOM ParisTech, France in 2017. His current research interests include person re-identification, image processing and deep learning technologies.

LI Dongfang received the M.S. degree in electronics and communications engineering from Harbin Engineering University, China in 2017. He has been engaged in deep learning technologies in ZTE Corporation since his graduation.

ZHENG Qingfang received the B.S. degree in civil engineering and computer applications from Shanghai Jiaotong University, China in 2002, and the Ph.D. degree in computer sciences from Chinese Academy of Sciences, China in 2008. He is currently the chief scientist of video technology in ZTE Corporation. His research interests include computer vision, video codec, video streaming and multimedia content analysis and retrieval. He has published around 10 papers in various journals and conferences.

JIA Xia received her B.S. degree and M.S. degree in control theory and control engineering from Taiyuan University of Technology and Dalian University of Technology, China in 1995 and 2001, respectively. She joined ZTE Corporation in 2001 and worked in the State Key Laboratory of Mobile Network and Mobile Multimedia Technology. Her main research interests include deep learning techniques, face detection and recognition, Re-ID, and activity detection and recognition.