Semantic-Guided Shared Feature Alignment for Occluded Person Re-IDentification

Xuena Ren
RENXUENA@IIE.AC.CN
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Dongming Zhang* ZHDM@CERT.ORG.CN and **Xiuguo Bao** BAOXIUGUO@139.COM
The National Computer Network Emergency Response Technical Team Coordination Center of China,Beijing, China

Editors: Sinno Jialin Pan and Masashi Sugiyama

Abstract

Occluded Person Re-ID is a challenging task under resolved. Instead of extracting features over the entire image which would easily cause mismatching, we propose Semantic-Guided Shared Feature Alignment (SGSFA) method to extract features focusing on the non-occluded parts. SGSFA parses human body regions through Semantic Guided Alignment(SGA) branch and aligns regions through Spatial Feature Alignment (SFA) branch simultaneously, and gets enriched representations over the regions for Re-ID. Dynamic classification loss of spatial features and their dynamical sequential combinations in the training stage help facilitate feature diversity. During the matching stage, we use only the visible feature shared by probe and gallery with no extra cues. The experiment results show that SGSFA achieves rank-1 of 62.3% and 50.5% respectively for Occluded-DukeMTMC and P-DukeMTMC-reID, surpassing the state-of-the-art by a large margin.

Keywords: Occluded person Re-ID,Semantic-Guided,Feature alignment,Dynamic classification loss.

1. Introduction

Person Re-IDentification (Re-ID) aims at retrieving the same person across different cameras with various backgrounds, poses, and positions. It has achieved great progress in recent years with the use of Convolutional Neural Networks (CNN). Most recent Re-ID models apply deep CNNs to extract features from the entire image, and then compute the similarity between query and gallery images. However, occluded person Re-ID is still a challenging task, because when a person is partially occluded, the representation extracted from the whole image might involve distractive information. It might lead to wrong retrieval results if a model does not differentiate the obstruction region and the person region. As shown in Fig. 1, when walking on public occasions, especially crowded places like subways, airports,

* Zhang Dongming is the corresponding author(zhdm@cert.org.cn).This work is supported in part by the National Key Research and Development Plan of China(2018YFB0804202) and in part by the National Natural Science Foundation of China(No.61672495 and No.U1736218).

or hospitals, a pedestrian is likely to be obstructed by other pedestrians and/or other objects like cars and bicycles.

Figure 1: Illustration of occluded images. The left part is from Occluded-REID, and the right part is from the P-DukeMTMC-reID. The occlusion position may be anywhere in the image, and may be top, bottom, left, or right. The occlusion area may be of any size. Coverings are also varied.

To tackle the challenges in Occluded Re-ID, PGFA Miao et al. (2019) applies pose estimator to generate the person landmark and gives higher weight to non-occluded person part, thus to suppress the noisy information from the occluded regions. However, pose estimator easily fails in some situations, which leads to PGFA's little lower performance than PCB Sun et al. (2018) on the holistic dataset. We follow the two principles proposed in PGFA Miao et al. (2019) :(1) In the feature construction stage, the model should pay more attention to the non-occluded parts. (2) In the matching stage, the global features should be divided into parts, and only the shared visible region between query and gallery images should be considered.

Following the two principles, different from PGFA, we propose to utilize semantic mask as guidance to align extracted features between gallery images and probe images, and name it "Semantic-Guided Shared Feature Alignment(SGSFA)". Compared with PGFA, the semantic branch provides a more reliable guidance to focus on the non-occluded regions and filter out the noise. Experiments on both the Occluded dataset and the holistic dataset demonstrate our method achieves competitive performance.

The main contributions of this paper are summarized as follows:

We propose a Semantic-Guided Shared Feature Alignment (SGSFA) CNN model for occluded Re-ID, which explicitly makes use of the learned visibility attention map to capture the emphasized human body parts in SGA branch and enriches the representation by incorporating spatial features of the SFA branch.

We make use of the shared part matching strategy for the aligned body feature and spatial feature according to the judgment of visibility based on our learned region attention

maps, without the extra computation caused by external human parsing or pose key-points detection model.

SGSFA achieves rank-1 of 62.3% and 50.5% on Occluded-DukeMTMC Miao et al. (2019) and P-DukeMTMC-reID Zhuo et al. (2018) benchmarks respectively, which improves over the state-of-the-arts by a large margin.

2. Related Work

Various methods have been proposed to solve the problem of Re-ID in the recent years Hermans et al. (2017), Chen et al. (2017), qi Li et al. (2017), Zheng et al. (2015), Sun et al. (2018). These methods can be grouped into feature extraction methods and metric learning strategies. These deep models, which focus on matching holistic images, are yet not invariable to uncontrollable variations like pose, low resolution, illumination, and occlusion. Among the factors, occlusion has been considered as the most challenging. The performances of existing Re-ID methods, which are based on global features, are degraded on occluded or partial data. Several methods have been proposed for processing non-full body images.

2.1. Partial Person Re-ID

Partial Re-ID Zheng et al. (2015b)aims to match partial probe image to gallery holistic images. Resizing model would normalize the partial image of a predefined size and then extract fixed-length image features, this method would cause deformation of the image and misalignment. Deep Spatial feature Reconstruction (DSR) He et al. (2018a) integrates sparse reconstruction in a deep model to reconstruct query from holistic gallery images, and reconstruction error is used to evaluate the matching of two arbitrary-sized feature maps. Spatial Feature Reconstruction (SFR) He et al. (2018b) combines spatial feature reconstruction with pyramid pooling to improve the model's robustness to scale, a foreground-background mask is used in FPR He et al. (2019) to avoid background and occlusion effects.

2.2. Occluded Person Re-ID

Occluded Re-ID task is to identify the same person from non-overlap cameras when given a detected target person with occlusions. It has attracted increasing attention due to its practical importance and it is more different to deal with for only partial body is available.

Artificial occlusion images are employed in model training and occlusion/non-occlusion binary classification (OBC) loss is used to distinguish the occluded image from the whole image in Zhuo et al. (2018). As its improvement, the teacher-student strategy trains a better model by using the saliency module to obtain distinctive featuresZhuo et al. (2019). PGFA Miao et al. (2019) applies pose estimator to generate the person landmark and pays more attention to non-occluded person parts, thus to suppress the noisy information from the occluded regions. HONet Wang et al. (2020) uses graph to learn higher-order relations and topological information. PVPM Gao et al. (2020) learns the distinguishing features through pose-guided attention, and self-mines the part visibility in an end-to-end framework. However, the success of such approaches heavily depends on the accuracy of pose estimators, and the use of pose estimator also causes extra computation cost.

3. Our Proposed Method

This section illustrates our proposed Semantic-Guided Shared Feature Alignment(SGSFA) method. It produces several pixel-wise attention maps, highlighting the visible body part while suppressing the occluded region or background. The proposed architecture is shown in Fig. 2. It comprises two main branches: a Semantic Guided Alignment(SGA) branch

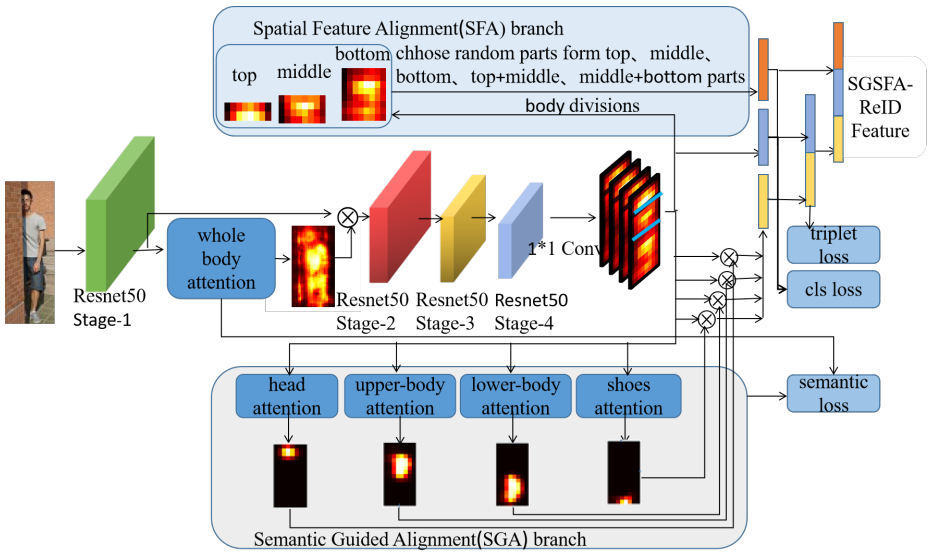

Figure 2: SGSFA uses ResNet-50 as backbone with some modifications. It contains a Semantic Guided Alignment(SGA) branch and a Spatial Feature Alignment(SFA) branch. In SGA, there are one whole body and four local part attentions. The four local attentions are head, upper body, lower body and shoes, respectively. These attention modules are guided by the corresponding body masks. In SFA, up, mid and low features are obtained through the division strategy of the proportion of human body structure. Features from SGA and SFA are concatenated together to form the final Re-ID features used in the inference stage.

that produces several pixel-wise attention maps and focuses on getting human features, and Spatial Feature Alignment(SFA) branch which generates spatially aligned features. The output is a feature vector,based on which the similarity between the query and gallery image is computed.

In our work, we adopt ResNet-50 as our backbone network and do a few modifications. The global average pooling (GAP) layer and the fully connected layer in the original network are removed. We set the stride of last residual block convolution operation to 1 so that the input image with a size of H×W will be 16 times down-sampled. A larger feature map with a size of H/16 × W/16 contains more spatial information. ResNet-50 has four groups of residual blocks,of which we impose the attention supervision on stage-1 and stage-4 respectively.

3.1. Semantic Guided Alignment Branch

The Semantic Guided Alignment Branch consists of whole-body attention and N body parts attention. As shown in Fig. 2. The body masks are used to guide the training of semantic attention and can pay more attention to the human body part other than the background, and discard the occluded region.

To extract global informative features and local discriminative features simultaneously, we consider 5 semantic (whole body, head, upper-body, lower-body, shoes) attention modules. It should be noted that different body attention is supervised in different convolution layers. The whole body's attention is inserted after stage-1, while the other four body part attentions are after stage-4. We use a CNN to create a global representation of input image x, denoted as $f(x)$. $M^{'}$ produced by our attention module is multiplied with the $f(x)$ to exclude background and occluded elements. Then the masked feature can be denoted as $f^{'}(x)$, then $f^{'}(x) = f(x) \cdot M^{'}$

For the whole body's attention, it takes the output from stage-1 of ResNet-50. The output of the whole-body attention is performed element-wise multiplication with the output features from stage-1. The reasons why we choose to design whole-body attention in stage-1 are as follows: (1) For the deep-based person recognition model, the person image feature usually is the output of the final fully-connected (FC) layer before the classification layer. However, each neuron in the fully connected layer aggregates the information of the previous network layer. In the final fully connected layer feature, the background or occlusion has been mixed into the human features. (2) CNNs encode more abstract and higher semantic-level features as they go deeper, and features obtained from the top fc layers are more likely sensitive to identities.

For the 4 local body part attention, each of them takes the output from stage-4 of the backbone ResNet-50, then carries out element-wise multiplication with the output features from stage-4, and finally outputs the local attentions. There are four semantic attentions, including head attention, upper-body attention,lower-body attention and shoes attention. The training of these four local attentions are guided by the masks generated by the off-the-shelf method, who is JPPNet Liang et al. (2018) in our experiments. The size of the output of local attentions is 16×8. The height of the map is 16 and the width is 8. Four local body part attentions are used to extract different semantic features. The first convolution features are shared by the four-body branches. Whole-body and local attention modules are shown in Fig.3 and Fig.4. A Convolution layer and a global pooling layer are followed. Each attention module outputs a $256d$ vector.

3.2. Spatial Feature Alignment Branch

The attention module guided by semantics can strictly locate head partition. However when pedestrians wear different clothes, the upper-body and lower-body partition may be labeled wrongly. E. g., when pedestrians wear dresses, the region of legs may be considered as upper body partition, and while a pedestrian wears jumpsuits, the region of upper body may be considered to be the lower body. Therefore, we consider part division according to the proportion of human body structure. Spatial features in SFA divided global feature into top, middle, and bottom parts according to the human body proportion, on which ergonomics tells us their respective proportions are 15%, 23%, and 62% respectively, i.e.,

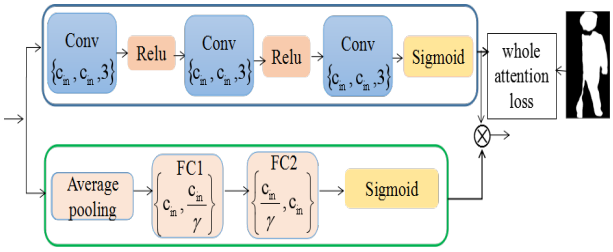

Figure 3: Structure of whole body attention module. Blue box is the spatial attention. Green box is the channel attention. γ is 8.

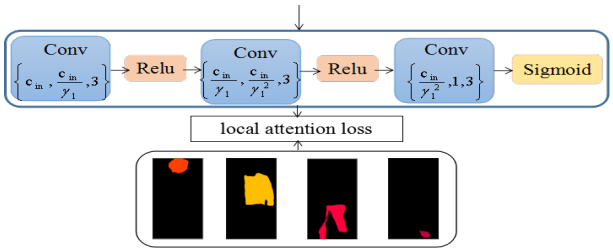

Figure 4: Structure of local body part attention module. γ_1 is 2.

starting from the position of the head, head accounts for 15% of the whole body, and the upper body accounts for 23%, and the lower body accounts for 62% of the human body. Ergonomics provides an efficient and reasonable way to divide the human body region.

3.2.1. Human body division

In order to efficiently align and segment according to the human body proportions, the position of the head needs to be determined first. The output of the head module in SGA branch is used to help determine where the head features start. It works as a learner as well as a supervisor. The learned head mask is scanned from top to bottom and from left to right, and then the first position with non-zero in the mask is used as the top position of the head. Then the other body partitions are labeled according to the above proportions.

3.2.2. Regularization with dynamic feature learning

The features have been spatially aligned, while detailed feature representations haven't been captured in all spatial regions. Further amplification of details is helpful to maintain robust model performance. To address this problem, we propose spatial regularization(SR) strategy to drive the network to learn discriminative representation for each local region. As the partitions are dynamically sampled in each batch, only limited number of regions are sampled to represent an identity. According to the visibility of each part, randomly selected local blocks f_{choose} participate in the training. Thus, the network is regularized to capture fine-grained details in dynamical region instead of fixed region(s). Herein,the

aligned spatial regions can be divided into five regions:

$$FR = \{top, middle, bottom, top + middle, middle + bottom\} \tag{1}$$

we denote the visibility of top, middle, and bottom partial blocks respectively with t, m, and b. If the values of their corresponding attention feature map are not all 0, the value of the label is assigned to 1, otherwise it is 0.

$$t, m, b \in \{0, 1\} \tag{2}$$

$$f_{choose} = \begin{cases} FR \ , if \ t = 1, m = 1, b = 1; \\ \{top, middle, top + middle\} \ , elseif \ t = 1, m = 1; \\ \{top\} \ , elseif \ t = 1; \\ \varnothing \ , otherwise \end{cases} \tag{3}$$

3.3. Feature Matching

The matching strategy is shown in Fig. 5. The final distance between query and gallery images consists of two parts. One is the distance of shared spatial part features and the other is the distance of shared visible semantic part features. In the inference stage, the output mask of the attention module is used to determine the visibility of the feature block. We calculate the average pixel value of m_i to obtain its visibility-label v_i:

$$v_i = \frac{\sum_{w=1}^{W} \sum_{h=1}^{H} m_i^{w,h}}{W \times H} \tag{4}$$

where W,H are the width and the height of the mask. If v_i exceeds a threshold, its corresponding feature is determined to be visible and the visible label can be defined as:

$$v_i = \begin{cases} 0, & v_i \leq \tau \\ 1, & v_i > \tau \end{cases} \tag{5}$$

For the positioning of the head, l_{head} is the maximum location of the learned head mask. Since we output a 16 * 8 feature map, if the entire head is visible, the proportion of the head is only 23%, so we choose line $l_{head} + 1$ as the starting position of the head feature.

Now the distance between query and gallery can be obtained by this:

$$d_i = D(f_i^q, f_i^g) \quad i = head, upper, lower, shoes, top, middle, bottom \tag{6}$$

$$dist = \sum v_i \cdot d_i \tag{7}$$

f_i^q and f_i^g are the i_{th} feature of the probe and gallery, d_i is the distance of i_{th} part. $\cdot$ denotes multiplication.

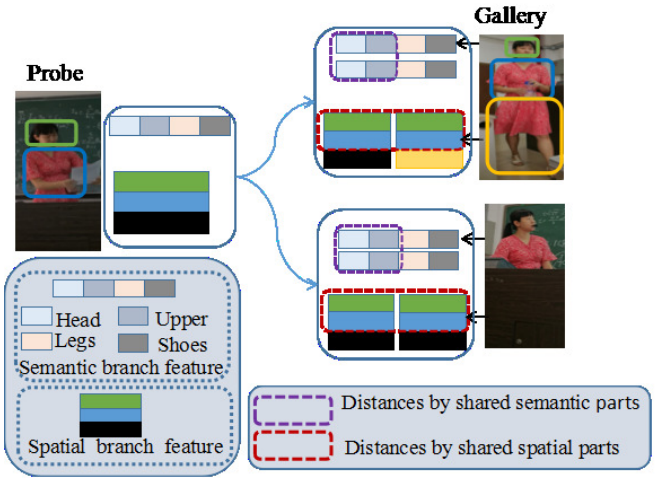

Figure 5: Matching strategy of SGSFA. The distance between probe and gallery images is measured by the shared semantic and spatial features.

3.4. Loss Function

In order to train our network, semantic attention loss is introduced.Three loss functions: semantic attention loss L_{se}, the batch hard triplet loss L_{tri}, dynamic classification loss are used to optimize our model.

Mask Attention Loss L: L on pixel-level supervision is formulated as a per-pixel binary cross-entropy loss (BCE loss):

$$L = BCELoss\left(M_s^{'}(x), M_s(x)\right) \tag{8}$$

where $M_s^{'}(x)$ are the predictions produced by spatial attention sub-module, and $M_s(x)$ represents the ground truth of pixel.

$$L_{se} = L_{wholebody} + L_{head} + L_{upper-body} + L_{lower-body} + L_{shoes} \tag{9}$$

Label smoothing (LS) Szegedy et al. (2016) is a widely used method to prevent over-fitting for a classification task. Label smooth changes the true probability distribution. By "softening" traditional one-hot type labels, it can effectively suppress over-fitting when calculating loss values. Thus we consider Cross-entropy with LSR as our classification loss, and it is computed as:

$$L_{cls} = \sum_{k=1}^{K} -y_k log\left(p_k\right) = -\left(1-\varepsilon+\frac{\varepsilon}{K}\right) logp_t - \frac{\varepsilon}{K}\sum_{i\neq t} logp_k \tag{10}$$

$$L_{loss} = \lambda(L_{tri} + L_{se}) + (1-\lambda)L_{d\_cls} \tag{11}$$

The total loss is a combination of these three types of loss.

4. Experiment

4.1. Datasets

Market-1501: Market-1501 contains 32668 images of 1501 identities. A total of six cameras were used, including 5 high-resolution cameras and one low-resolution camera. 12936 images of 751 identities are divided into training set. 19732 images of the remaining 750 identities are divided into testing set. 3368 images are in query set. The maximum number of query images is 6 for an identity.

DukeMTMC-reID: DukeMTMC-reID is a subset of the DukeMTMC dataset for image-based Re-IDentification. It consists of 36411 images of 1812 identities from 8 different cameras. 1404 identities are appearing in more than two cameras and 408 identities (distractor ID) appearing in only one camera. 16522 images of 702 persons are divided into training set. 19889 images of the remaining identities are divided into testing set, with 2228 in query set and 17661 in gallery set.

Partial-REID Zheng et al. (2015b) is the first partial person Re-ID dataset including 900 images of 60 persons. Each person has 5 full-body person images, 5 partial person images and 5 occluded person images with various occlusions. All the images were collected at a campus.

Occluded-REID Zhuo et al. (2018) is an occluded person dataset captured by mobile cameras with different viewpoints and backgrounds. There are two folders, the occluded one and the whole one, including 2000 images of 200 identities. Each identity has 5 full-body person images and 5 occluded person images with different types of occlusions.

Occluded-DukeMTMC Miao et al. (2019) contains 15,618 training images, 17,661 gallery images, and 2,210 occluded query images. Occluded-DukeMTMC contains 9%/100%/10% occluded images. All probe images are occluded and the query contains both holistic and occluded images.

P-DukeMTMC-reID Zhuo et al. (2018) is a modified version based on DukeMTMC-reID dataset. There are 12,927 images (665 identifies) in training set, 2,163 images (634 identities) for querying and 9,053 images in the gallery set.

4.2. Implement Details

We choose ResNet-50 pre-trained on ImageNet Deng et al. (2009) as our CNN backbone by removing the fully connected layer.The stride of last residual module convolution operation of ResNet50 is set to be 1, so that high-resolution features and more spatial information can be obtained.For classifiers in SA branch,the feature maps from the last residual block go through a Batch Normalization layer Ioffe and Szegedy (2015) and a fully connect layer followed by a softmax function. Horizontal flipping, normalization and random erasing Zhong et al. (2017),which are commonly used data augmentation strategies in Re-ID,are applied in our methodology. In the process of erasing, the pixel value of randomly selected area is the mean value of the whole image.The semantic maps are using off-the shelf deep methods Gong et al. (2017). The whole body masks size are resized to 64×32, while the local body masks size is 16×8. We divide the 20 semantic categories into 4 groups.

4.2.1. Training Details

The input images are resized to 256×128.The margin of batch hard triplet loss is 0.3. The batch size is 64. There are 16 identities in each mini-batch and 4 images for each identity. The initial learning rate for the backbone network is 0.0003, decreases by 0.1 at 50, 80, 100,120 epochs. Our model is totally trained on 120 epochs.The model is implemented on the Pytorch platform and trained with one NVIDIA 1080Ti GPU.

4.2.2. Evaluation Metrics

We adopt standard metrics that are currently used in most literature, namely Cumulative Matching Characteristic (CMC) and mean average precision (mAP) to evaluate our method. rank-1, rank-5, rank-10 and mAP results are reported. All results reported in this paper are under single query setting.

4.3. Experimental Results

We make extensive experiments on different datasets.

4.3.1. Results on Occluded Datasets

Results on Occluded-DukeMTMC: Occluded images are in both the probe and gallery. We conduct experiment and compare the performances with the exsiting methods on Occluded-DukeMTMC. The results are listed in Table 1. As can be seen that, SGSFA gets the top performance among the compared approaches,and obtains 62.3% /47.4% in rank-1/mAP. SGSFA surpasses HONet Wang et al. (2020) by +7.2% rank-1 accuracy and +3.6% mAP, which is a large margin. From Table 1, we can find that the occluded methods always perform better than those holistic approaches by a large margin. This result validates that it is essential to propose a specifically designed framework for the occluded Re-ID task. Besides,our proposed SGSFA achieves comparable results of 64.1%, 68.3% and 50.5% at rank-1 on the Occluded ReID, Partial-REID and P-DukeMTMC-reID dataset. We think that the improvements come from the followed aspects: 1) part matching strategy is more suitable for occluded Re-ID task than global feature learning 2) shared visible parts matching works better than using all parts 3) body proportion division strategy is more likely to align features than horizontal partitioning.

For the large-scale dataset P-DukeMTMC-reID, we further run experiments to evaluate the performance when optimizing the model with the target training set.The results are shown in Table 2. SGSFA achieves 86.4% at rank-1. This indicates the superiority of our model under the supervised setting for occluded person Re-ID.

4.3.2. Results on Holistic Datasets

Although the performance of current methods for occlusion has been greatly improved, some strategies cannot achieve satisfactory results on the holistic dataset. The performance cannot be balanced between the full body and the occlusion dataset.Thus,We also apply our method on holistic person Re-ID datasets, Market-1501 and DukeMTMC-reID. As shown in Table 6, our method achieves comparable performances with state of-the-art on both datasets, which indicates the good generality of our method.

Table 1: Performance comparison on Occluded-DukeMTMC.

Method	rank1	rank5	rank10	mAP
LOMO+XQDA Liao et al. (2014)	8.1	17.0	22.0	5.0
DIM Yu et al. (2017)	21.5	36.1	42.8	14.4
Part Aligned Zhao et al. (2017)	28.8	44.6	51.0	20.2
Random Erasing Zhong et al. (2017)	40.5	59.6	66.8	30.0
Adver Occluded Huang et al. (2018)	44.5	-	-	32.2
HACNN Li et al. (2018)	34.4	51.9	59.4	26.0
Part Bilinear Suh et al. (2018)	36.9	-	-	-
FD-GAN Ge et al. (2018)	40.8	-	-	-
PCB Sun et al. (2018)	42.6	57.1	62.9	33.7
DSR He et al. (2018a)	40.8	58.2	65.2	30.4
SFR He et al. (2018b)	42.3	60.3	67.3	32.0
FPR He et al. (2019)	-	-	-	-
PGFA Miao et al. (2019)	51.4	68.6	74.9	37.3
HONet Wang et al. (2020)	55.1	-	-	43.8
SGSFA	62.3	77.3	82.7	47.4

Table 2: Performance on the P-DukeMTMC-reID dataset .

Model	rank-1	rank-5	rank-10	mAP
Performance under unsupervised setting				
IDE Zheng et al. (2015a)	36.0	49.3	55.2	19.7
PCB Sun et al. (2018)	43.6	57.1	63.3	24.7
Teacher-S Zhuo et al. (2019)	18.8	24.2	32.2	22.4
PVPM Gao et al. (2020)	50.1	63.0	68.6	29.4
SGSFA	50.5	64.4	69.5	28.5
Performance under supervised setting				
IDE Zheng et al. (2015a)	82.9	89.4	91.5	65.9
PCB Sun et al. (2018)	79.4	87.1	90.0	63.9
Teacher-S Zhuo et al. (2019)	51.4	50.9	-	-
PVPM Gao et al. (2020)	85.1	91.3	93.3	69.9
SGSFA	86.4	91.5	92.7	66.3

5. Analyses

5.1. The contribution of whole attention

We have carefully made a set of comparative experiments to verify the influence of the whole body's attention module. The results are shown in Table 5. SGSFA achieves better performance than SGSFA $_{w/o\ whole\ attention}$ on both rank-1 accuracy and mAP. Without the whole attention module, the highly abstract convolutional layer features in stage-4 will be contaminated by the occlusion, resulting in impure learning features. This module acts as a filter for occlusion and background.

Table 3: Performance comparisons with the holistic and unsupervised occluded methods on small Partial-REID dataset. The $1^{st}/2^{nd}$ best results are in red and blue.

Method	Partial-REID			
	rank-1	rank-5	rank-10	mAP
IDE Zheng et al. (2015a)	51.7	69.0	80.3	52.4
HACNN Li et al. (2018)	37.0	64.0	75.3	40.4
Bilinear Suh et al. (2018)	57.7	77.3	85.7	59.3
PCB Sun et al. (2018)	66.3	84.0	91.0	63.8
PCB+RPP Sun et al. (2018)	63.7	82.3	90.0	61.2
Teacher-S Zhuo et al. (2019)	69.2	76.6	85.8	73.1
PGFA Miao et al. (2019)	68.0	82.0	86.7	56.2
PVPM Gao et al. (2020)	75.3	88.7	92.3	71.4
SGSFA	68.2	82.5	87.7	65.3

Table 4: Performance comparisons with the holistic and unsupervised occluded methods on Occluded-REID datasets. The $1^{st}/2^{nd}$ best results are in red and blue.

Method	Occluded-REID			
	rank-1	rank-5	rank-10	mAP
IDE Zheng et al. (2015a)	52.6	68.7	76.6	46.4
HACNN Li et al. (2018)	29.1	44.7	54.7	26.1
Bilinear Suh et al. (2018)	54.9	70.8	77.7	50.3
PCB Sun et al. (2018)	59.3	75.2	83.2	53.2
PCB+RPP Sun et al. (2018)	55.8	74.4	81.2	51.3
Teacher-S Zhuo et al. (2019)	55.0	64.5	77.3	59.8
PGFA Miao et al. (2019)	57.1	77.9	84.0	56.2
PVPM Gao et al. (2020)	66.8	82.0	88.4	59.5
SGSFA	63.1	77.8	84.1	53.2

5.2. Trade-off Coefficient between SGA and SFA Branch

To evaluate the impact of the two branches: Semantic Guided Feature Alignment Branch and Spatial Feature Alignment Branch, we conduct many experiments with different trade-off coefficients of λ, which is introduced in equation 11, on Occluded-DukeMTMC. When λ is 1, only the semantic feature is considered,while when λ is 0, only spatial features take effect. The experiment results with different λ are shown in Fig. 6. As can be seen that, the performance of SGSFA is gradually improving when λ increases from 0 to 0.2, which demonstrates the effectiveness of the SFA. This result also proves that the semantic branch plays a major role. If we continue to increase λ, the SGSFA will not improve performance further, which indicates that SGSFA is sensitive to the λ in this range. On the experiments, we conclude that 0.8 is best choice for λ. However, we believe that more flexible trade-off between the two branches is expected in the future.

Table 5: Performance comparison of whole attention module. SGSFA $_{w/o\ whole\ attention}$ denotes our method without the whole attention module in the semantic branch.

Method	rank1	rank5	rank10	mAP
SGSFA $_{w/o\ whole\ attention}$	57.3	73.6	78.8	41.2
SGSFA	63.4	78.0	83.3	48.2

Table 6: Performance comparisons on holistic datasets

Method	Market-1501		DukeMTMC-reID	
	rank-1	mAP	rank-1	mAP
BoW+kissme Zheng et al. (2015)	44.4	20.8	25.1	12.2
PAN Zheng et al. (2017)	82.8	63.4	71.7	51.5
DSR He et al. (2018a)	83.5	64.2	-	-
MultiLoss qi Li et al. (2017)	83.9	64.4	-	-
TripletLoss Hermans et al. (2017)	84.9	69.1	-	-
Adver Occluded Huang et al. (2018)	86.5	78.3	79.1	62.1
MultiScale Chen et al. (2017)	88.9	73.1	79.2	60.6
SSP-ReID Quispe and Pedrini (2019)	89.3	75.9	80.1	66.1
MLFN Chang et al. (2018)	90.0	74.3	81.0	62.8
HA-CNN Li et al. (2018)	91.2	75.7	80.5	63.8
PGFA Miao et al. (2019)	91.2	76.8	82.6	65.5
PCB Sun et al. (2018)	92.4	77.3	81.9	65.3
SPreID Kalayeh et al. (2018)	92.5	81.3	85.9	73.3
AlignedReID Zhang et al. (2017)	92.6	82.3	-	-
SGSFA	92.3	80.2	84.7	70.8

5.3. Comparison with PCB and PGFA

SGSFA prevails PCB and PGFA due to the following facts. PCB Sun et al. (2018) method can learn local information well, but it cannot avoid the influence of occlusion. Secondly, the part features learned by PCB may be spatially misaligned. PGFA Miao et al. (2019) employs pose estimator to detect key point information to obtain non-occluded pedestrian

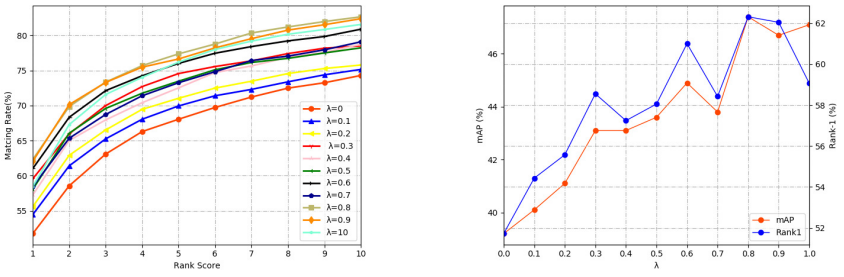

Figure 6: Performances with different λ

features, but in the feature matching stage, distance by pose-guided global feature is not completely shared by probe and gallery, which leads to improper feature matching between partial and holistic. This is also the cause of mismatching.

6. Conclusion

In this paper, we propose a Semantic-Guided Shared Feature Alignment(SGSFA) method to tackle the occluded person Re-ID problem. Our method can suppress occluded or contaminated information by through the semantic attention modules. The SR strategy exploits the discriminative details in each rigid parts by learning with dynamical representation. Besides, SGSFA utilizes the semantic features in semantic branch and spatial features in the spatial branch that are shared between the gallery and probe images for matching. Experiments show that SGSFA achieves state-of-the-art performance on Occluded-DukeMTMC. This illustrates that aligning the features is key to occluded preson Re-ID. In the future, we will further improve the balancing the different branches to facilitate Re-ID in real complex scene.

References

X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018.

Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2590–2600, 2017.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.

Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. *ArXiv*, abs/2004.00230, 2020.

Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018.

K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6757–6765, July 2017. doi: 10.1109/CVPR.2017.715.

Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7073–7082, 2018a.

Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018b.

Lingxiao He, Yinggang Wang, Wu Liu, Xingyu Liao, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8449–8458, 2019.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *ArXiv*, abs/1703.07737, 2017.

H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang. Adversarially occluded samples for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018.

Shengcai Liao, Yang Hu, and Stan Z. Li. Joint dimension reduction and metric learning for person re-identification. *CoRR*, abs/1406.4216, 2014. URL http://arxiv.org/abs/1406.4216.

Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 542–551, 2019.

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

Rodolfo Quispe and Helio Pedrini. Improved person re-identification based on saliency and semantic parsing with deep neural network models. *Image and Vision Computing*, 92: 103809, 2019.

Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. *ArXiv*, abs/1804.07094, 2018.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian-Tao Sun. High-order information matters: Learning relation and topology for occluded person re-identification. *ArXiv*, abs/2003.08177, 2020.

Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *CoRR*, abs/1711.08106, 2017. URL http://arxiv.org/abs/1711.08106.

Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3219–3228, 2017.

L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015a.

Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015b.

Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017. URL http://arxiv.org/abs/1708.04896.

Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. 2018.

Jiaxuan Zhuo, Jianhuang Lai, and Peijia Chen. A novel teacher-student learning framework for occluded person re-identification. *CoRR*, abs/1907.03253, 2019. URL http://arxiv.org/abs/1907.03253.