Pose-Guided Feature Alignment for Occluded Person Re-Identification

Jiaxu Miao[1,2] Yu Wu[1,2] Ping Liu[2] Yuhang Ding[1] Yi Yang[2†]
[1]Baidu Research [2] ReLER, University of Technology Sydney
{jiaxu.miao,yu.wu-3}@student.uts.edu.au,dyh.ustc.uts@gmail.com, {ping.liu,yi.yang}@uts.edu.au

Abstract

*Persons are often occluded by various obstacles in person retrieval scenarios. Previous person re-identification (re-id) methods, either overlook this issue or resolve it based on an extreme assumption. To alleviate the occlusion problem, we propose to detect the occluded regions, and explicitly exclude those regions during feature generation and matching. In this paper, we introduce a novel method named **Pose-Guided Feature Alignment** (PGFA), exploiting pose landmarks to disentangle the useful information from the occlusion noise. During the feature constructing stage, our method utilizes human landmarks to generate attention maps. The generated attention maps indicate if a specific body part is occluded and guide our model to attend to the non-occluded regions. During matching, we explicitly partition the global feature into parts and use the pose landmarks to indicate which partial features belonging to the target person. Only the visible regions are utilized for the retrieval. Besides, we construct a large-scale dataset for the Occluded Person Re-ID problem, namely **Occluded-DukeMTMC**, which is by far the largest dataset for the Occlusion Person Re-ID. Extensive experiments are conducted on our constructed occluded re-id dataset, two partial re-id datasets, and two commonly used holistic re-id datasets. Our method largely outperforms existing person re-id methods on three occlusion datasets, while remains top performance on two holistic datasets.*

1. Introduction

Person re-identification (re-id), who aims at retrieving a probe person image from a collection of images, has achieved a great progress in recent years [31, 13, 32, 38, 24, 42, 16, 4]. Generally, previous person re-id approaches extract features from whole images and utilize those features as visual representations to match gallery candidates. To construct an effective representation, previous methods either directly utilize global person features [31, 13] or combine local features of body parts [32, 38, 24, 42].

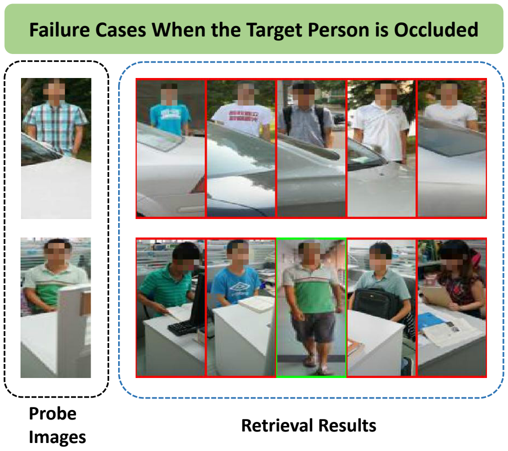

Figure 1: Some failure retrieval examples of previous re-id approaches [32] when the target person is occluded. Green and red rectangles indicate correct and error retrieval results, respectively. When an occlusion exists, previous methods fail to achieve a satisfactory result since it does not suppress the noise introduced by the occlusions.

However, the methods proposed in those previous works [31, 13, 32, 38, 24, 42] did not consider a situation that the target person is occluded by various obstructions like cars, trees, or other people. 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. For instance, as illustrated in Fig.1, given a query image of a person occluded by a white car, previous methods may retrieve wrong person images with a similar car by mistake.

Recently, some works [45, 11, 12] attempt to solve the occlusion issue. As shown in the first row of Fig. 2, in their **Partial Re-ID** problem setting, the probe images are occluded by obstructions while the gallery images are still holistic. To depress the unexpected information introduced by the occlusion, they first manually crop the occluded target person in probe images and then use the non-occluded

[†]Corresponding author.
[‡]Part of this work was done when Jiaxu Miao and Yu Wu were interns at Baidu Research.

parts as the new query. However, there are two limitations in the Partial Re-ID problem setting: (1) They require a strong assumption that all gallery images are holistic, which does not hold all the time. (2) They need a manually cropping operation, which is inefficient considering the huge size of gallery set if the gallery set contains occluded images too. In addition, the manual process might introduce human bias to the cropped results.

Different from the Partial Re-ID problem, we propose the **Occluded Re-ID** problem, in which both probe and gallery images contain occlusions. All probe images are occluded in the Occluded Re-ID problem, making at least one occluded image exist when retrieving. In addition to the holistic images, gallery set also contains occluded images, which is consistent with the real world scenarios. Besides, the Occluded Re-ID problem does not employ manually cropping process considering efficiency and avoiding human bias. The difference between the Partial Re-ID problem and our Occluded Re-ID problem is shown in Fig. 2. To facilitate the research on the Occluded Re-ID problem, we introduce a large-scale dataset named **Occluded-DukeMTMC**, derived from DukeMTMC-reID [46, 28]. In the new dataset, all query images are occluded by large variety of occlusion (*e.g.*, trees, cars, other persons), while gallery images contain both holistic and occluded images. Details about Occluded-DukeMTMC will be introduced in Section 3.

To tackle this more challenging Occluded Re-ID problem, we propose two strategies to disentangle the information of visible regions from occluded regions: (1) In the feature constructing stage, the model should pay more attention to the non-occluded parts. (2) In the matching stage, we explicitly partition the global features into parts and only consider the shared visible region between probe and gallery images.

Motivated by these two strategies, we propose to utilize pose landmarks as guidance to align extracted features between gallery images and probe images, and name it "Pose-Guided Feature Alignment (**PGFA**)". Compared with previous works, our proposed PGFA has several advantages. First, unlike [45, 11, 12], PGFA does not need any manually cropping operation and is more efficient. Second, the meta information of the detected landmarks can explicitly guide the model to focus on the non-occluded person regions and filter out the information of occluded regions in both feature constructing and matching stage.

Experiments on the Occluded-DukeMTMC dataset show that our method outperforms previous works [32, 11, 12] by a large margin. On two Partial Re-ID datasets and two commonly used holistic benchmarks, our approach still achieves a competitive performance. The main contributions of this paper are summarized as follows:

- We introduce a challenging large-scale occluded re-id dataset, Occluded-DukeMTMC, which is by far the largest dataset focusing on Occluded Person Re-ID.
- We propose PGFA, an effective method for the Occluded Re-ID problem. PGFA takes advantages of the information from detected human landmarks, and deploy them as guidance to attend to non-occluded region during representation construction stage and align partial representations during matching stage.

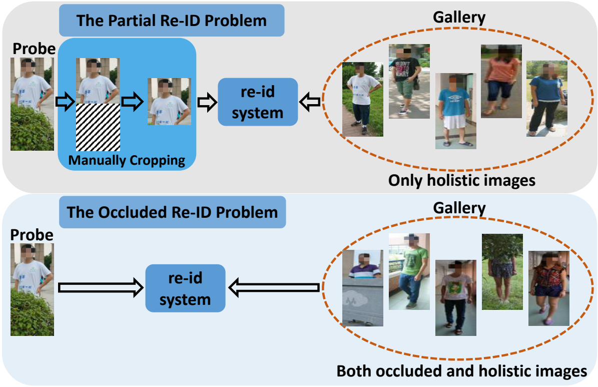

Figure 2: Previous *Partial Re-ID* problem (1st row) and our *Occluded Re-ID* problem (2nd row). The Partial Re-ID problem assumes that probe images are occluded and all gallery images are holistic. In this setting, each probe image is manually cropped, and its visible part is used to query. The Occluded Re-ID problem assumes both the probe and gallery contain occlusions. Besides, the new setting does not employ manually cropping process.

2. Related Work

2.1. Deep Person Re-ID

Recent years, deep learning methods [39, 34, 13, 27, 3, 25, 7, 6, 36, 21] for person re-id show significant superiority on retrieval accuracy. Recent works [15, 42, 24, 32] utilize deeply part-based features learning method and further improve the state-of-the-art performance on the holistic person re-id problem. For example, Kalayeh *et al.* [15] extract several region parts with human parsing methods, learn the feature for each piece, and assemble final discriminative representations with part-level features. Sun *et al.* [32] uniformly partition the feature map and learn each part-level feature by multiple classifiers. Zhao *et al.* [42] and Liu *et al.* [24] extract part-level features by attention-based methods. All these approaches [15, 42, 24, 32] achieve better performance by assembling part-level features.

However, as shown in Fig. 1, when facing the aforementioned Occluded Re-ID problem, those previous works mix information of the target person and obstacles into the final feature representation and usually fail in real scenarios. A recent work [49] make attempts to solve occlusion problems in person re-id. However, they ignore the situation that both query and gallery contain occlusions.

Dataset	Training Set		Gallery		Query	
	#ID	#Image	#ID	#Image	#ID	#Image
Occ-DukeMTMC	702	15,618	1,110	17,661	519	2,210
Partial-REID [45]	-	-	60	300	60	300
Partial-iLIDS [44]	-	-	119	119	119	119

Table 1: The scale of three person re-id datasets focusing on occlusion problems. Two (Partial-REID [45] and Partial-iLIDS [44]) for Partial Re-ID, one (Occ-DukeMTMC) for Occluded Re-ID.

2.2. Partial Person Re-ID

Occlusion occurs in real-world scenarios where only partial regions of the target person are available for person re-id. Several partial re-id methods have been proposed to tackle the Partial Re-ID problem. Zheng *et al*. [45] propose a local patch-level matching model called Ambiguity-sensitive Matching Classifier (AMC) and introduce a global part-based matching model called Sliding Window Matching (SWM). He *et al*. [11] propose an alignment-free method called Deep Spatial Feature Reconstruction (DSR). DSR sparsely reconstructs the spatial probe maps from spatial maps of gallery images which are faster than the SWM method. He *et al*. [12] further utilize a dictionary learning-based Spatial Feature Reconstruction (SFR) to match different sized feature maps for the Partial Re-ID problem.

In [45, 11, 12], each probe image is occluded while the gallery images are holistic. To deal with the occlusion, [45, 11, 12] manually crop probe images and utilize the visible parts as new probe images. The manual cropping is not efficient and might introduce human bias to jeopardize the final performance. Different from these works, our work considers a more general situation that both probe and gallery images have occlusions on target person, and does not need a manually cropping process in it.

2.3. Pose-guided Person Re-ID

Pose landmarks indicate the body position of persons and are conductive to various vision problems. Recently, some person re-id methods [9, 33, 29, 23, 30] employ pose information to facilitate person re-id models. These methods utilize pose landmarks to generate person images [9, 23] or align body parts [30, 29]. These pose-driven methods focus on tackling the large variations introduced by human pose while our method uses pose information to tackle the occlusion problems. Besides, we utilize pose landmarks to encode the position information of body parts and use shared-region part-to-part comparison between query and gallery images. Based on the experimental results, all of those specific design in our method are proved effective and efficient to solve the Occluded Person Re-ID problem. Zhang *et al*. [41] also utilize the pose-guided attention mechanism to tackle the occlusion problems. However, they focus on the detection task while ours focuses on the retrieval problem (re-id).

3. The Occluded-DukeMTMC Dataset

To facilitate the research on the Occluded Person Re-ID problem, we introduce Occluded-DukeMTMC, a large-scale occluded person re-identification dataset, derived from the DukeMTMC-reID dataset [46, 28].

3.1. Properties of Occluded-DukeMTMC

There are a few properties to make Occluded-DukeMTMC appealing. First, it is the largest occluded person re-id dataset to date. The **training set** of Occluded-DukeMTMC contains $15,618$ images covering 702 identities in total. The **testing set** contains $1,110$ identities, including $17,661$ gallery images and $2,210$ query images. As shown in Table 1, the size of Occluded-DukeMTMC is ten times larger than the size of the previous occlusion dataset [45, 44]. Second, previous datasets, either do not focus on solving the occlusion problems [43], or depend on a strong assumption: only probe images are occluded [45, 44]. Compared to those datasets, Occluded-DukeMTMC is more difficult and practical since both probe and gallery images have occlusions. Third, there are rich variations in Occluded-DukeMTMC, including different viewpoints and a large variety of obstacles, including cars, bicycles, trees, other persons *et al*.

3.2. Data Collection

In the original DukeMTMC-reID, the training, query, and gallery set contain 14%/15%/10% occluded images, respectively. Apparently, the original dataset is not applicable to evaluate occluded person re-id approaches due to its small occluded sample numbers.

We manually re-split DukeMTMC-reID so that our Occluded-DukeMTMC contains 9%/**100%**/10% occluded images. All probe images are **occluded** by manually selecting from both the gallery[1] and query set of the original dataset. Therefore, there always exists at least one occluded image when calculating the pairwise distance between query and gallery images. When we select query images, the image which contains more than one person or a person occluded by obstacles such as trees or cars is annotated as an occluded image.

When constructing the training set, we manually removed 934 images from original DukeMTMC-reID training set, because these 934 images contain exactly the same obstacles as in the testing set. Those images may lead the model to "remember" these specific occlusion patterns in the test set, which jeopardizes the generality of the trained model.

[1] Re-id evaluation does not count the images in the same camera so there is no worry to query the same image in the gallery.

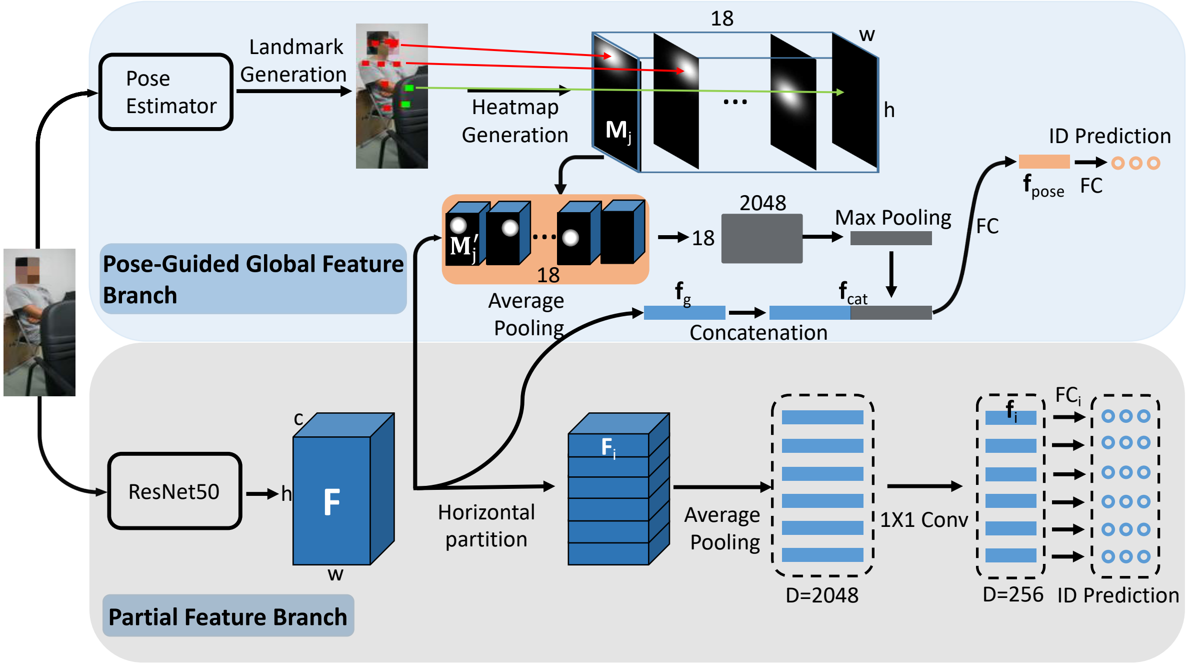

Figure 3: Pipeline of the Pose-Guided Feature Alignment method. The red points in the person image represent the visible landmarks, and the green point represents the invisible landmark. The pipeline has two branches. In the partial feature branch, the global feature map is horizontally partitioned as partial features. Then we use an average pooling layer and convolutional layer to reduce the dimensions of the partial feature maps, followed by multiple fully connected layers to predict the ID of each input image. In the pose-guided global feature branch, heatmaps are generated by pose landmarks and multiply the global feature map element-wisely. Then we concatenate the masked feature and the average pooled global feature, and reduce the feature dimension to generate the pose-guided global feature.

4. Methodology

This section illustrates our proposed Pose-Guided Feature Alignment (PGFA) method, consisting of a representation construction stage and a matching stage.

4.1. Representation Construction

The architecture for representation construction is shown in Fig.3. As illustrated in Fig.3, it is a two-branch architecture. One is Partial Feature Branch, the other one is Pose-Guided Global Feature Branch, which utilizes a pose estimator to detect human landmarks and to guide robust representation construction.

Following [32], PGFA uses ResNet50 [10] without the average pooling layer and fully connected layer as the backbone to extract global feature maps from given images. Motivated by [32, 38], in PGFA, the stride of conv4_1 is set to 1. As a result, after passing through our backbone network, an input image with a size of $H \times W$ will output feature maps with a spatial dimension of $H/16 \times W/16$, which is larger than that $(H/32 \times W/32)$ in the original ResNet50 [10]. A larger feature map is useful in our case because information of the target person and occlusions is easier to disentangle in a broader spatial dimension.

Formally, we denote the feature map extracted from ResNet50 backbone as $\mathbf{F} \in \mathbb{R}^{h \times w \times c}$, in which h, w, c denotes the height, width and channel number, respectively. Then the feature map $\mathbf{F}$ is input to two branches: the **partial feature branch** and the **pose-guided global feature branch**.

4.1.1 Partial Feature Branch

In the partial feature branch, $\mathbf{F}$ is horizontally partitioned into p parts, denoted as $\mathbf{F}_i$, $i = 1, ..., p$. Then each part of feature map $\mathbf{F}_i$ is processed by average pooling and 1×1 convolutional layers to reduce dimension from $2,048$ to 256, obtaining the partial feature vector $\mathbf{f}_i$. Each partial feature $\mathbf{f}_i$ is fed into a corresponding fully connected layer FC_i and a softmax layer is employed to predict the ID of each input image. The output for each given image I in Partial Feature Branch is a set of predictions $\hat{y}_i$, where $i = 1, ..., p$. Denote the loss of the partial feature branch as L_{part},

$$L_{part} = \sum_{i=1}^{p} CE(\hat{y}_i, y)\,, \tag{1}$$

where CE denotes the cross-entropy loss, p is the part number, $\hat{y}_i$ is the probability of predictions, and y is the ground truth ID. This branch is similar to [32], and readers are encouraged to refer it.

4.1.2 Pose-Guided Global Feature Branch

Pose Estimation. PGFA employs a human pose estimator to detect human landmarks from person images. The estimator is pre-trained on the COCO dataset [20]. Given an input image, PGFA utilizes the pose estimator to extract N landmarks for the inside person, where N is 18 in this paper. For each landmark, the pose estimator predicts its coordinates and confidence score. We set a threshold γ to filter out the landmarks with too small confidence scores, which are smaller than γ.[2] Formally, the landmarks is

$$LM_j = \begin{cases} (cx_j, cy_j) & \text{if } S_j^{conf} \geq \gamma \\ 0 & \text{else} \end{cases} \quad (j = 1, ..., N)\,, \quad (2)$$

where LM_j denotes the jth landmark and cx_j,cy_j denote the coordinates of the jth landmark, S_j^{conf} is the confidence score and γ is the threshold.

Similar to [26], landmarks are utilized to generate heatmaps consisting of a 2D Gaussian centered on the ground truth location. When $LM_j = 0$, the value of the corresponding heatmap is set to 0. The generated heatmap is denoted as $\mathbf{M}_j$, $j = 1, ..., N$. Each heatmap is downsampled to the size of $h \times w$ by bilinear interpolation.

Pose-Guided Global Feature Construction. The pose-guided global feature branch aims to integrate the global feature maps information and the pose information from the target person. As shown in Fig.3, firstly the feature map $\mathbf{F}$ is average pooled as the global feature, $\mathbf{f}_g$. Then the heatmap $\mathbf{M}_j$, $j = 1, ..., N$, multiply the feature map $\mathbf{F}$ element-wisely and output the pose-guided feature map $\mathbf{M}'_j$, $j = 1, ..., N$. Since each heatmap $\mathbf{M}_j$ has explicitly encoded the information of different regions on the target person, i.e., which region is occluded, the pose-guided feature maps $\mathbf{M}'_j$ can focus on non-occluded parts of the target person and depress the information from occluded regions.

Each guided feature map $\mathbf{M}'_j$ will pass through an average pooling layer to produce a 2048-D feature vector, which corresponds to the region containing the specific landmarks. Finally, PGFA performs the max pooling operation over all feature vectors and concatenate them with the global feature $\mathbf{f}_g$. The concatenated feature is denoted as $\mathbf{f}_{cat}$, as shown in Fig.3. The max pooling operation makes the feature vector fuse information of visible body parts, ignore the occluded parts and redundant partial information. $\mathbf{f}_{cat}$ is fed into a fully connected layer to reduce the dimension from 4,096 to 256, denoted as the pose-guided global feature $\mathbf{f}_{pose}$. A softmax layer is employed to predict the ID of each input image, and the output of Pose-guided Global Feature Branch is the prediction $\hat{y}$ of each input image I. We utilize the cross-entropy loss, which is denoted as L_{pose}, for the pose-guided global branch.

$$L_{pose} = CE(\hat{y}, y)\,, \quad (3)$$

where CE denotes the cross-entropy loss, $\hat{y}$ denotes the prediction and y denotes the ground truth ID.

The final loss function is

$$Loss = \lambda L_{part} + (1 - \lambda) L_{pose}\,, \quad (4)$$

where λ is a coefficient to balance the contributions from the part feature branch and the pose-guided feature branch, while L_{part} is the softmax cross entropy loss for the partial feature branch and L_{pose} is the softmax cross entropy loss for the pose-guided global feature branch.

4.2. Representation Matching

The matching strategy is shown in Fig. 4. The final distance between query and gallery images consists of two parts. One is the distance of partial features in the shared visible region and the other is the distance of the pose-guided global feature. Since the confidence score of detected landmarks can indicate which part of the target person is occluded and which part is not, they can guide us to obtain part labels. Specifically, for each part $i = 1, ..., p$, its part label $l_i \in \{0, 1\}$ is illustrated as follows:

$$l_i = \begin{cases} 1 & \text{if } \exists cy_j \in [\frac{i-1}{p}H, \frac{i}{p}H) \\ 0 & \text{else} \end{cases} \quad (j = 1, ..., N)\,, \quad (5)$$

where cy_j denotes the jth longitudinal coordinate of landmark LM_j, i denotes the ith part of an image and H denotes the height of the image.

Now the distance measure function d_i of the ith part is illustrated as follows:

$$d_i = D(\mathbf{f}_i^q, \mathbf{f}_i^g) \quad (i = 1, ..., p)\,, \quad (6)$$

where d_i denotes the distance calculated by the ith partial feature, $D()$ denotes the distance metric, which is the cosine distance in our paper. $\mathbf{f}_i^q$, $\mathbf{f}_i^g$ denote the ith partial feature of the query and gallery image, respectively.

Denote the distance calculated by the pose-guided global feature as d_{pose},

$$d_{pose} = D(\mathbf{f}_{pose}^q, \mathbf{f}_{pose}^g)\,, \quad (7)$$

where $D()$ denotes the cosine distance metric, $\mathbf{f}_{pose}^q$, $\mathbf{f}_{pose}^g$ denote the pose-guided global feature of the query and gallery image, respectively. The final distance

$$dist = \frac{\sum_{i=1}^{p} (l_i^q \cdot l_i^g) d_i + d_{pose}}{\sum_{i=1}^{p} l_i^q \cdot l_i^g + 1}\,, \quad (8)$$

where d_i denotes the distance calculated by the ith partial feature, d_{pose} denotes the cosine distance calculated by the

[2] In the situation that a person is occluded by another person, we compare the number of landmarks of each person and assume that the person with the largest number of landmarks is the target person while others are obstacles.

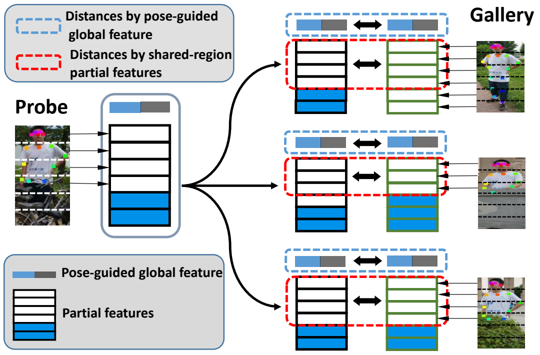
Distances by pose-guided global feature
Distances by shared-region partial features
Probe
Gallery
Pose-guided global feature
Partial features

Figure 4: Matching strategy of PGFA. The distance between probe and gallery images is measured by using the pose-guided global feature and the partial features in the shared visible region.

pose-guided global feature, p denotes the part number and $\cdot$ denotes multiplication. l_i^q and l_i^g denote the ith part label of the query and gallery image, respectively. If the ith partial feature of both the query and gallery image contain visible parts, $l_i^q \cdot l_i^g = 1$. Else, $l_i^q \cdot l_i^g = 0$. The larger the $dist$ is, the higher similarity between the probe and gallery images.

5. Experiments

5.1. Datasets and Evaluation Measures

To demonstrate the efficacy of our method on the Occluded Re-ID problem, we evaluate the proposed PGFA model on our constructed Occluded-DukeMTMC and two Partial Re-ID datasets: Partial-REID [45], Partial-iLIDS [44]. We also evaluate our method on two holistic datasets: Market-1501 [43], DukeMTMC-reID [46, 28].

Occluded-DukeMTMC contains 15,618 training images, 17,661 gallery images, and 2,210 occluded query images. The experiment results on Occluded-DukeMTMC will demonstrate the superiority of our method in Occluded Person Re-ID problems, let alone that our method does not need any manually cropping procedure as pre-process.

Partial-REID [45] is a specially designed partial person dataset that includes 600 images from 60 people, with five full-body images and five partial images per person. These images are collected at a university campus from different viewpoints, backgrounds, and different types of severe occlusion. All probe images are occluded person images, while all gallery images are holistic images.

Partial-iLIDS [44] is a simulated partial re-id dataset based on the iLIDS dataset [44]. The iLIDS dataset contains a total of 476 images of 119 people captured by multiple non-overlapping cameras. We conduct experiments on Partial-REID and Partial-iLIDS to demonstrate the effectiveness of our proposed method when facing the Partial Re-ID problem.[3]

Market-1501 [43] is used to verify our method on the holistic re-id situation. Market-1501 contains few of occluded person images and can be treated as a holistic re-id dataset. Market-1501 consists of $32,668$ images of $1,501$ subjects captured by six cameras.

DukeMTMC-reID [46, 28] contains 1,404 identities, 16,522 training images, 2,228 queries, and 17,661 gallery images. Although there exist occluded person images, the holistic images in DukeMTMC-reID are much more than the occluded ones, so that this dataset can be treated as a holistic re-id dataset in previous works [46, 17, 32].

Evaluation Metrics. We use Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP) to evaluate the quality of different person re-identification models. All the experiments are performed in a single query setting.

5.2. Implementation Details

We use ResNet50 [10] as our backbone and make a minor modification on it: removing the average pooling layer and fully connected layer, setting the stride of conv4_1 to 1. We initialize our model by the ImageNet [5] pre-trained model. In our experiment setting, the input image is resized to 384×128 and augmented by random flipping and random erasing [48]. We set the batch size to 32 and the training epoch number to 60. On Occluded-DukeMTMC, Market-1501 [43], and DukeMTMC-reID [46, 28], the base learning rate is initialized at 0.1 and decayed to 0.01 after 40 epochs, the coefficient λ is set to 0.2. On Partial-REID and Partial-iLIDS, the base learning rate is initialized at 0.02, the coefficient λ is set to 0.9.

To detect landmarks from occluded images, we use AlphaPose [8, 37] pre-trained on the COCO dataset [20]. We keep the landmarks whose confidence scores are larger than 0.2.

5.3. Results Comparison

Results on Occluded-DukeMTMC. Table.2 shows the result of our method and previous works. The approaches in the first group are designed for the holistic person re-id problem. The methods in the second group employ pose estimation methods. The methods in the third group are designed for the Partial Re-ID problem. Our PGFA achieves 51.4% Rank-1 accuracy and 37.3% mAP, which outperforms all the previous methods. Compared to the strong competing method PCB [32], PGFA surpasses it by +8.8% Rank-1 accuracy and +3.6% mAP. This is because PGFA explicitly utilizes pose information to depress the noisy information from the occluded regions.

In the fourth group, $\text{PGFA}_{\text{w/o partial}}$ denotes using only the pose-guided global feature for matching, while $\text{PGFA}_{\text{w/o global}}$ denotes utilizing only the partial features for matching. From the table, we can find that combining two

[3] Unlike previous works [45, 11, 12], our method does not need manually cropping on Partial-REID and Partial-iLIDS.

Method	Rank-1	Rank-5	Rank-10	mAP
LOMO+XQDA [18]	8.1	17.0	22.0	5.0
DIM [40]	21.5	36.1	42.8	14.4
Part Aligned [42]	28.8	44.6	51.0	20.2
Random Erasing [48]	40.5	59.6	66.8	30.0
HACNN [17]	34.4	51.9	59.4	26.0
Adver Occluded [14]	44.5	-	-	32.2
PCB [32]	42.6	57.1	62.9	33.7
Part Bilinear [30]	36.9	-	-	-
FD-GAN [9]	40.8	-	-	-
DSR [11]	40.8	58.2	65.2	30.4
SFR [12]	42.3	60.3	67.3	32.0
$\text{PGFA}_{\text{w/o partial}}$(ours)	42.5	60.2	67.3	30.4
$\text{PGFA}_{\text{w/o global}}$(ours)	46.0	65.4	72.0	34.4
PGFA(ours)	**51.4**	**68.6**	**74.9**	**37.3**

Table 2: Performance comparison on Occluded-DukeMTMC. The methods in the 1st group are designed for the holistic re-id problem. The methods in the 2nd group utilize the pose information. The methods in the 3rd group are designed for the Partial Re-ID problem. The 4th group is our methods.

Method	Partial-REID		Partial_iLIDS	
	Rank-1	Rank-3	Rank-1	Rank-3
MTRC [19]	23.7	27.3	17.7	26.1
AMC+SWM [45]	37.3	46.0	21.0	32.8
DSR [11]	50.7	70.0	58.8	67.2
SFR [12]	56.9	78.5	63.9	74.8
PGFA(ours)	**68.0**	**80.0**	**69.1**	**80.9**

Table 3: Performance comparison on Partial-REID and Partial-iLIDS.

branches achieves better performance than using either one branch alone.

Results on Partial-REID and Partial-iLIDS. We compare the results on Partial-REID and Partial-iLIDS with several existing partial person re-id methods, including MTRC [19], AMC+SWM [45], DSR [11], and SFR [12]. Note that partial person re-id approaches above utilize the manually cropped probe images, while our method needs no additional manually cropping process. Same as previous works [11, 12], we train our model using the Market-1501 training set. In practice, we set the part number p to 6 in both training procedure and testing procedure and achieves the best performance. As shown in Table.3, the Rank-1/Rank-3 of our methods PGFA achieve 68.0%/80.0% and 69.1%/80.9% on Partial-REID and Partial-iLIDS, respectively, which outperforms all the previous partial person re-id approaches [19, 45, 11, 12]. Compared to the strongest competing method SFR [12], PGFA surpasses it by +11.1% Rank-1 accuracy on Partial-REID, while surpasses it by +5.2% Rank-1 accuracy on Partial-iLIDS, which is a large margin.

Results on Market-1501 and DukeMTMC-reID. We also apply our method on holistic person re-id datasets, Market-1501 and DukeMTMC-reID. As shown in Table.4, our method achieves comparable performances with state-of-the-art on both datasets, which indicates the good generality of our method.

Method	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
BoW+kissme [43]	44.4	20.8	25.1	12.2
SVDNet [31]	82.3	62.1	76.7	56.8
PAN [46]	82.8	63.4	71.7	51.5
PAR [42]	81.0	63.4	-	-
Pedestrian[47]	82.0	63.0	-	-
DSR [11]	83.5	64.2	-	-
MultiLoss [35]	83.9	64.4	-	-
TripletLoss [13]	84.9	69.1	-	-
Adver Occluded [14]	86.5	78.3	79.1	62.1
APR [22]	87.0	66.9	73.9	55.6
MultiScale [4]	88.9	73.1	79.2	60.6
MLFN [2]	90.0	74.3	81.0	62.8
PCB [32]	92.4	77.3	81.9	65.3
PGFA(ours)	91.2	76.8	82.6	65.5

Table 4: Performance comparison on the holistic re-id datasets Market-1501 and DukeMTMC-reID.

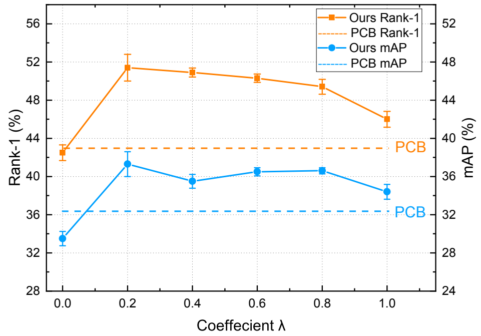

Figure 5: Varying Trade-off Coefficients λ. The dash lines in the figure denote the performance of PCB [32]. When λ is 0, we use only the pose-guided global feature. When λ is 1, which means only the partial features take effect, the difference between PCB and ours is that we use the shared region matching strategy and our PGFA outperforms PCB by +3.4% Rank-1 accuracy.

5.4. Ablation Study

We conduct extensive ablation studies on Occluded-DukeMTMC to analyze each component of PGFA.

Varying Trade-off Coefficients. To evaluate the impact of the two branches: Partial Feature Branch and Pose-Guided Global Feature Branch, we conduct a test with different trade-off coefficients of λ, which is defined in Equation. 4, and report the Rank-1 accuracy and mAP in Fig. 5. We increase λ from 0 to 1. When λ is 0, only the pose-guided global feature branch takes effect, which achieves

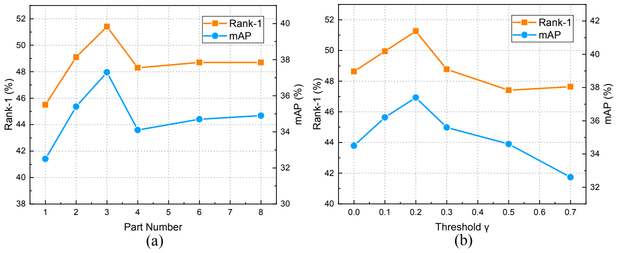

Figure 6: The impact of the part number p **(a)** and the threshold γ of pose estimation results **(b)**.

42.5% Rank-1 and 30.4% mAP. When λ is 1, only the partial feature branch takes effect and the Rank-1 accuracy is 46.0%, which surpasses PCB by $+3.4\%$. This is because the shared-region matching strategy in PGFA helps to filter the noise from obstacles, while PCB fails to tell which region is occluded but has to utilize all information, even when some of them are extracted from obstacles. When λ is between 0 and 1, both the partial feature branch and the pose-guided global feature branch are taking effect, which achieves much better performance than only using one of the two branches. When λ is set between 0.2 and 0.8, the performance does not change dramatically, which indicates that PGFA is not sensitive to the λ in this value range.

The Sensitivity of the Part Number p. The number of parts p determines the granularity of the part feature. When $p = 1$, the learned feature is a global feature. As illustrated in Fig.6(a), the performance when $p = 1$ is always poorer than the performance when $p > 1$, which proves the necessity of introducing partition strategy in extracted features. As p increase to 3, PGFA achieves the best performance: 51.4% Rank-1 accuracy and 37.2% mAP. When p is larger than 3, the performance starts to decrease slowly. This is because that when the part number p is too large, some non-occluded parts might not contain any landmarks, and will be filtered out in matching since their corresponding l_i is 0 in Equation. 5.

The Impact of the Threshold γ. As defined in Equation. 2, γ is the threshold to filter the landmarks whose confidence score is too small. As shown in Fig.6(b), when γ is too small or too large, the performance is poor. This is because when γ is too small (for example, 0), all detected landmarks are chosen. Therefore, the information from the occluded regions will be used for the representation construction and matching. This will inevitably bring noisy information and deteriorate the final performance when there are occlusions in probe and gallery images. When γ is too large, many landmarks will be discarded. The corresponding regions of those discarded landmarks, although they might do not have any occlusions, are unnecessarily thrown away.

The Impact of the Pose Estimation Algorithm. We test two different pose estimation algorithms, AlphaPose [8], and OpenPose [1] in PGFA. The results are shown in Table.5. From the table, we can find that OpenPose achieves similar performance with AlphaPose.

Method		Rank-1	Rank-5	mAP
AlphaPose [8]	PGFA$_{\text{w/o partial}}$	42.5	60.2	30.4
	PGFA$_{\text{w/o global}}$	46.0	65.4	34.4
	PGFA	51.4	68.6	37.2
OpenPose [1]	PGFA$_{\text{w/o partial}}$	42.4	59.3	29.7
	PGFA$_{\text{w/o global}}$	46.2	65.4	33.3
	PGFA	49.1	66.7	35.3

Table 5: Performance comparison of pose estimation algorithms. PGFA$_{\text{w/o partial}}$ denotes our method without the partial feature branch. PGFA$_{\text{w/o global}}$ denotes our method without the pose-guided global feature branch.

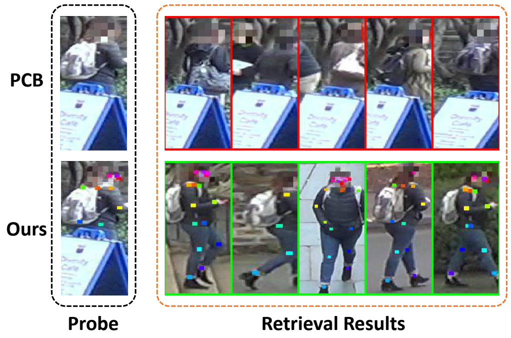

Figure 7: Comparison of the PCB method and our PGFA method. Green and red rectangles indicate correct and error retrieval results, respectively.

5.5. Visualization

Fig. 7 shows some retrieval examples of the PCB [32] method and our PGFA method on Occluded-DukeMTMC. The retrieval results show that PCB is prone to mix the information of the target person and obstacles, resulting in retrieving a wrong person with a similar obstacle. On the contrary, our PGFA can work successfully in the same situation.

6. Conclusion

In this paper, we make contributions to tackle the Occluded Person Re-ID Problem. First, we propose the PGFA method, which outperforms existing approaches on the Occluded Re-ID problem. By taking advantage of information from detected landmarks, our method can suppress the noisy information from the occluded regions on the target person. Besides, PGFA utilizes the partial features in the shared region between the gallery and probe images for matching. Second, to facilitate the research about the Occluded Re-ID problem, we introduce a large-scale dataset, Occluded-DukeMTMC.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[2] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018.

[3] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017.

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *ICCVW*, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. *arXiv preprint arXiv:1907.04194*, 2019.

[7] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *TOMCCAP*, 2018.

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[9] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NIPS*, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, 2018.

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

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[14] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *CVPR*, 2018.

[15] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gkmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[17] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[18] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[19] Shengcai Liao, Anil K Jain, and Stan Z Li. Partial face recognition: Alignment-free approach. *TPAMI*, 2013.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[21] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019.

[22] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019.

[23] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, 2018.

[24] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017.

[25] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019.

[26] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[27] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, 2017.

[28] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCVW*, 2016.

[29] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[30] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.

[31] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.

[32] 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 *ECCV*, 2018.

[33] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, 2017.

[34] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

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

[36] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018.

[37] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018.

[38] Fu Yang, Wei Yunchao, Zhou Yuqian, Shi Honghui, Huang Gao, Wang Xinchao, Yao Zhiqiang, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019.

[39] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[40] 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. *arXiv preprint arXiv:1711.08106*, 2017.

[41] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *CVPR*, 2018.

[42] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[44] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.

[45] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, 2015.

[46] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[47] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *TCSVT*, 2018.

[48] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

[49] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *ICME*, 2018.