Pose-guided Visible Part Matching for Occluded Person ReID

Shang Gao[1], Jingya Wang[3], Huchuan Lu[1,2*], Zimo Liu[1]
[1]Dalian University of Technology, [2]Pengcheng Lab
[3]UBTECH Sydney AI Center, The University of Sydney
gs940601k@gmail.com, jingya.wang@sydney.edu.au, lhchuan@dlut.edu.cn, lzm920316@gmail.com

Abstract

Occluded person re-identification is a challenging task as the appearance varies substantially with various obstacles, especially in the crowd scenario. To address this issue, we propose a Pose-guided Visible Part Matching (PVPM) method that jointly learns the discriminative features with pose-guided attention and self-mines the part visibility in an end-to-end framework. Specifically, the proposed PVPM includes two key components: 1) pose-guided attention (PGA) method for part feature pooling that exploits more discriminative local features; 2) pose-guided visibility predictor (PVP) that estimates whether a part suffers the occlusion or not. As there are no ground truth training annotations for the occluded part, we turn to utilize the characteristic of part correspondence in positive pairs and self-mining the correspondence scores via graph matching. The generated correspondence scores are then utilized as pseudo-labels for visibility predictor (PVP). Experimental results on three reported occluded benchmarks show that the proposed method achieves competitive performance to state-of-the-art methods. The source codes are available at `https://github.com/hh23333/PVPM`

1. Introduction

Person re-identification (ReID) aims to retrieve a probe pedestrian from non-overlapping camera views. It is an important research topic in computer vision field with various applications, such as autonomous driving, video surveillance and activity analysis [26, 19, 11]. Most existing ReID approaches design the matching model with the assumption that the entire body of the pedestrian is available. However, this assumption is hard to be satisfied due to the inevitable occlusions in real-world scenarios. For example, as shown in Figure. 1, a person may be occluded by other pedestrians, static obstacles like trees, walls and cars, etc. Therefore, it is essential to seek an effective method to solve this occluded person re-identification problem.

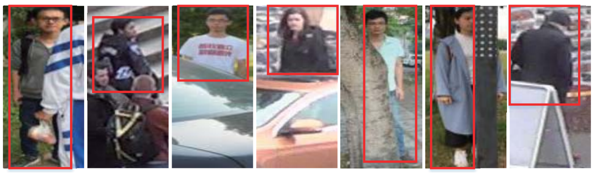

Figure 1. Illustration of occluded person re-id. Red bounding box indicates the target person which is occluded by diversified obstacles with different colors, sizes and positions.

There are two main challenges for the occluded person ReID task. First, the global image-based supervision for conventional person ReID may involve not only the information of the target person but also the interference of occlusion. The diversified occlusions, such as colors, positions and sizes, enhance the difficulty of getting a robust feature for the target person. Second, the occluded body parts sometimes show more discriminative information while the non-occluded body parts share a similar appearance, leading to the problem of mismatching.

An intuitive solution is to detect the non-occluded body parts and then match the correspondents separately. As there is no ground truth annotation for the occluded part, most existing methods directly utilize visibility cues from other tasks with different data source, e.g. body mask [1] and pose landmark estimation [16], but suffering a huge data bias without the flexibility on target domain. In this work, we proposed a Pose-guided Visible Part Matching (PVPM) network by directly mining the visible score in a self-learning manner. The concept of the proposed approach is demonstrated in Figure 2. As shown, PVPM includes two main components: a pose-guided part attention (PGA) network and a pose-guided visibility predictor (PVP) in an end-to-end framework. The training of the part visibility predictor is supervised by a pseudo-label obtained by solving a feature correspondence problem via graph matching. In the end, the final score can be computed via the summation of body-part distance aggregation weighted by the visibility score.

In conclusion, the main contribution of the proposed

[*]Corresponding Author

method is as following:

- We propose a Pose-guided Visible Part Matching (PVPM) method that jointly learn the discriminative features with pose-guided attention and predict the part visibility in an end-to-end framework.
- We train the visibility prediction model under a self-supervised manner, of which its pseudo-label generating process is regrad as a feature correspondence problem and is solved via graph matching.
- The proposed approach achieves superior performance on multiple occlusion datasets including Partial-REID [31], Occluded-REID [36] and P-DukeMTMC-reID [36].

2. Related Work

Occluded Person ReID. Most existing ReID works [9, 27, 14, 35, 4, 10, 30] focus on training the model without taking the occlusions into considerations. However, the occlusion can not be ignored especially in the crowd scenes like airports or hospitals. To address this problem, Zhou *et al.* [36] propose multi-task losses that force the network to distinguish between simulated occluded samples and non-occluded samples, so as to learn a robust feature representation against occlusion. Besides, a co-saliency network[37] is proposed to train model paying attention to the person body parts. More recently, Miao *et al.* [16] utilize pose landmarks to disentangle the useful information from the occlusion noise. Although the improvement has been made by introducing the pose landmarks, its untrainable pose-guided region extracting and the predefined landmarks visibility still limit the matching performance. Instead of simply using predefined regions and part visibility that learn from other data sources with limited flexibility and data bias, we try to self-mine the part visibility from target data and adapt pose-guide attention accordingly in a unified framework.

Part-based Person ReID. Part-based person ReID approaches exploit local descriptors from different regions to enhance the discriminative ability and robustness of the algorithm. A straightforward way to do this is to slice the person images or feature maps into uniform partitions [27, 24]. In [24], Sun *et al.* partition feature maps into p horizontal stripe and train each part embedding with non-shared classifiers. One can also extract the local features by pose-driven RoI extraction [28, 20], human parsing results [9] or learning attention regions based on appearance feature [10, 29, 15] or pose feature [22]. For example, Zhao *et al.* [28] propose to utilize pose detection results to generate local region by a manual-designed cropping manner, and then fuse the part features gradually. Kalayeh *et al.* [9] utilize human semantic parsing results to extract body part features. Suh *et al.* [22] propose to generate part maps from prior pose information and then aggregate all parts with a bilinear pooling. In [10, 29, 15], they attempt to use appearance-based attention maps to exploit local information. Although the local features are considered in model design, there are no cues for partial occlusion, leading mismatch in the complex environment.

Self-Supervised Learning. For the specific task of part visibility prediction, it appears that the precise label of each body part is unavailable to be obtained. This motivated us to solve this problem under the self-supervised learning manner. Self-supervised learning is proposed to learn feature from unlabelled data by introducing a so-called pretext task, for which a target objective can be computed with self-generated pseudo-label, such as spatial and temporal structure [8, 17, 25], or context similarity [18, 3]. Noroozi *et al.* [17] define the pretext task as recognizing the order of the shuffled patches from an image. Caron *et al.* [3] treat the cluster assignments as pseudo-labels to learn the parameters of ConvNet. Unlike the pretext task designed above, in this work, we generate the pseudo-label for visibility predictor by self-mine which utilizes the characteristic of part correspondence in positive pairs via graph matching.

3. Pose-Guide Visible Part Matching

In this work, we present a pose-guided visible part matching framework which aggregates local features with the visible scores to solve the mismatching problem for occluded person ReID task. To better understand the proposed method, we demonstrate the pipeline and the training process in Figure 2 and Algorithm Box 1, with the related notations illustrated in Table 1. This framework includes a pose encoder (PE), a pose-guide attention mask generator (PGA), a pose-guided visibility score predictor (PVP) and a feature correspondence model for generating the pseudo-label for training PVP. In Sec. 3.1 and 3.2, we introduce the methodology details of the PGA and PVP modules. In Sec. 3.3, we claim the strategy about how to obtain the pseudo-label of part correspondence to supervise the training of PVP. Last, in Sec. 3.4, we demonstrate the formulation of the loss functions we employed in this method.

3.1. Part Features with Pose-Guide Attention

Obviously, discriminative part features play an important role for the circumstance when the target person facing with occlusions. This motivates us to get the body part features by fusing the appearance features with pose-guided attention maps. With a given pedestrian image I, we first extract the appearance feature maps $F \in \mathbb{R}^{C \times H \times W}$ via a CNN backbone network, where C, H, W denote the number of pixel in the channel, height, and width dimensions of feature maps, respectively.

The pose-guided attention mechanism consists of three components: pose estimation, pose encoder and part atten-

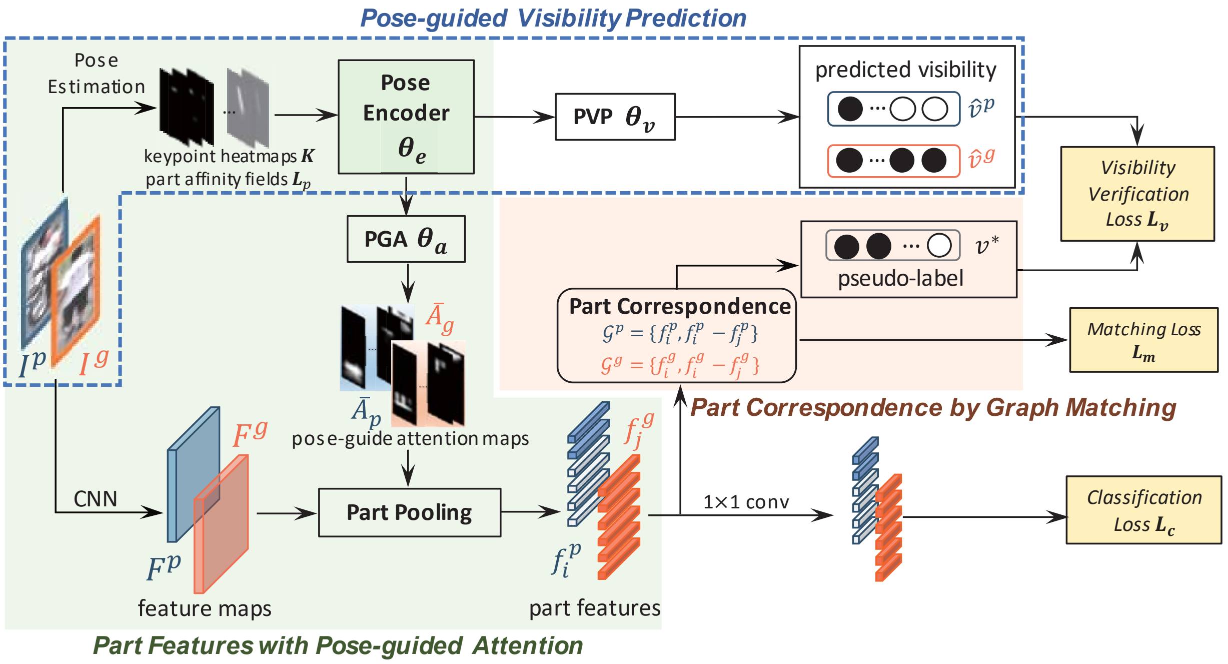
Pose-guided Visibility Prediction
Pose Estimation
keypoint heatmaps K
part affinity fields L_p
Pose Encoder θ_e
PVP θ_v
predicted visibility
$\hat{v}^p$
$\hat{v}^g$
Visibility Verification Loss L_v
PGA θ_a
$\bar{A}_g$
$\bar{A}_p$
pose-guide attention maps
v^*
pseudo-label
Part Correspondence
$\mathcal{G}^p = \{f_i^p, f_i^p - f_j^p\}$
$\mathcal{G}^g = \{f_i^g, f_i^g - f_j^g\}$
Matching Loss L_m
Part Correspondence by Graph Matching
I^p
I^g
CNN
F^g
F^p
feature maps
Part Pooling
f_j^g
f_i^p
part features
1×1 conv
Classification Loss L_c
Part Features with Pose-guided Attention

Figure 2. The pipeline of the proposed PVPM approach. It consists of three key components: a pose-guided attention (PGA) model for part feature pooling, a pose-guided visibility predictor (PVP), and a feature correspondence model for providing pseudo-label for the training of PVP. Three loss functions are employed, including L_v, L_m and L_c.

tion generator. We employ the Openpose [2] method for pose estimation to extract the key point heatmaps K and the part affinity fields L_p of each input image. The pose encoder then takes $P = K \oplus L_p$ as input and embeds the pose information into a high-level pose feature $F_{pose} = PE(P;\theta_e)$. For the part attention generator which focuses on a specific body part, a 1×1 Convolutional layer and a following Sigmoid function is adopted on pose features F_{pose} to estimate a stack of 2-dimensional maps A, each element $a_i^{h,w}$ in A indicates the degree that the location (h, w) from feature maps F lies in the i-th part:

$$A = PGA(F_{pose};\theta_a) \in \mathcal{R}^{N_p \times H \times W} \tag{1}$$

where N_p is the number of pre-defined parts, θ_a is the parameters of the convolutional layer. Furthermore, we hope the network could focus on the nonoverlapping region so that each part could extract complementary features which are more discriminative and robust when fusing them all. Thus, we only maintain the maximum activation along the first channel for each part map, which is formulated by,

$$\bar{A}_i = A_i \odot [\arg\max_i A_i]|_{onehot}^{C} \tag{2}$$

$\odot$ is the Hadamard Product, $[\arg\max_i A_i]|_{onehot}^{C}$ means to get the index of maximum value along the channel dimension and turn it into a one-hot vector at each spatial location.

In the end, the i-th part feature f_i can thus be obtained via a part weighted pooling, which is formulated by,

$$f_i = \frac{1}{\|\bar{A}_i\|} \sum_{h=1}^{H} \sum_{w=1}^{W} \bar{a}_i^{h,w} \odot F^{h,w} \tag{3}$$

$$\|\bar{A}_i\| = \sum_{h=1}^{H} \sum_{w=1}^{W} \bar{a}_i^{h,w} \tag{4}$$

where $F^{h,w}$ is the column vector of F at position (h, w), $\bar{a}_i^{h,w}$ denotes the element lies in the location (h,w) of $\bar{A}_i$.

3.2. Pose-Guide Visibility Prediction

After representing the pedestrian using part-based features, an intuitive way to calculate the distance is to compute the global part-to-part distances. However, for occluded ReID, some patches appear in one view may not be exposed in other views. Therefore, a reasonable way is to only establish the correspondence between simultaneously visible parts and compute the distance accordingly. We propose to utilize a pose-guide visibility score predictor (PVP) to estimate the visibility for each part.

We implement the PVP method via a four-layer tiny network which consists of a global average pooling layer, a Convolutional layer of 1×1 filter, a BatchNorm layer and a Sigmoid activation layer. With an input pose feature F_{pose}, we can predict the visibility score through,

$$\hat{v} = PVP(F_{pose};\theta_v) \in \mathcal{R}^{N_p} \tag{5}$$

Table 1. Notation definition.

I^p, I^g	probe/gallery images
F^p, F^g	probe/gallery feature extracted by CNN
F_{pose}	pose features after pose encoder
$\tilde{A}^p, \tilde{A}^g$	pose-guide attention maps of probe/gallery
f_i^p, f_j^g	i/j-th part features of probe/gallery
$\mathcal{G}^p, \mathcal{G}^g, M$	graph of probe/gallery and its affinity matrix
$\hat{v}^p, \hat{v}^g$	predicted visibility of probe/gallery via PVP
v^*	visibility optimized by correspondence learning
$\theta_e, \theta_v, \theta_a$	parameter of pose encoder, PVP, PGA
λ	regularization coefficient
T	maximal iteration for training
N_p	parts number for each image

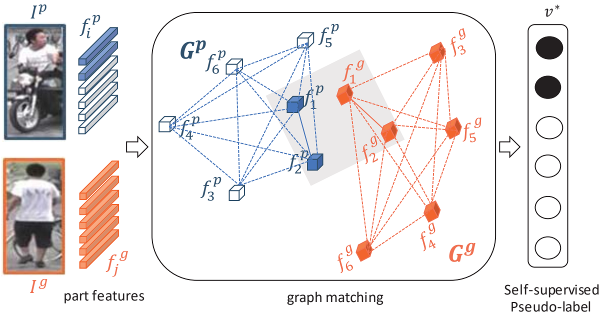

Figure 3. Pseudo-label estimation via graph matching.

When it comes to the testing stage, given a probe image I_p and a gallery image I_g, the distance considering the visibility between them can be calculated as:

$$d = \frac{\sum_{i=1}^{N_p} \hat{v}_i^p \hat{v}_i^g d_i}{\sum_{i=1}^{N_p} \hat{v}_i^p \hat{v}_i^g} \tag{6}$$

where d_i is the cosine distance of the i-th part features, $\hat{v}_p^i$ and $\hat{v}_g^i$ denote the visibility score of the i-th part of the probe image and gallery image, respectively.

3.3. Pseudo-Label Estimation by Graph Matching

The ground-truth visibility label of each part is usually unavailable. This motivates us to seek a method that can automatically reveal the visible part without further requirement of manually occlusion annotation in a self-supervised way. For a given positive image pair I^p, I^g, we observe that (1) the relevance of a part-pair appears to be high only when both parts in I^p, I^g are visible. (2) the relevance between the edges of two visible parts within the image will also be highly correlated.

Based on these observations, instead of training $\hat{v}$ directly, we train the product of part visibility scores of positive pairs to approximate their correspondence. Thus, we consider the pseudo-label generation process as a part feature correspondence problem which can be solved by graph matching. For better understanding, we illustrate an example of how to obtain the pseudo-label between two input images in Figure 3.

Specifically, for a given positive pair, we represent them via two graphs $\mathcal{G}^p = (\mathcal{V}^p, \mathcal{E}^p)$ and $\mathcal{G}^g = (\mathcal{V}^g, \mathcal{E}^g)$, where each element $\mathcal{V}_i$ and $\mathcal{E}_{i,j}$ denote the parts(nodes) features f_i and edges features $\{f_i - f_j\}$ respectively. In our task, only one-to-one matching between corresponding nodes of two graphs are adopted. A binary indicator vector $v \in \{0,1\}^{N_P}$ is employed to represent the correspondence of the two input parts from $\mathcal{G}^p$ and $\mathcal{G}^g$, where v_i set as 1 if the i-th part pair is selected for matching, otherwise 0. The affinity matrix M is conducted with the relational similarity values between edges and nodes where the inner product is used to calculate similarity. Specifically, we encode the compatibility of corresponding two nodes in the diagonal M_{ii} as:

$$M_{i,i} = \langle f_i^p, f_i^g \rangle \tag{7}$$

and encode the compatibility of corresponding two edges features in the non-diagonal component M_{ij} as:

$$M_{i,j} = \langle \frac{\mathcal{E}_{i,j}^p}{\|\mathcal{E}_{i,j}^p\|_2}, \frac{\mathcal{E}_{i,j}^g}{\|\mathcal{E}_{i,j}^g\|_2} \rangle - \hat{M}_{i,j} \tag{8}$$

where $\hat{M}_{i,j}$ is the moving average of $M_{i,j}$.

Same as another graph matching method [21], we model graph matching as an Integer Quadratic Programming (IPQ) problem and incorporate a regularization term on the number of activated nodes:

$$\arg\max_{v} \quad v^T M v - \bar{\lambda}^T v \quad \text{s.t. } v \in \{0,1\}^{N_p}. \tag{9}$$

$$\bar{\lambda} = \lambda \hat{M}_{diag} \tag{10}$$

where λ is a balanced parameter and $\hat{M}_{diag}$ is the moving average of diagonal components of M. We set $\bar{\lambda}$ to be proportional to the moving average of parts similarity to make it more adaptive to data as well as narrow down the scope of hyper-parameter selection. By optimizing Eq.(9), we can obtain the optimal solution v^* which indicates which part pair is appropriate to be matched. Then it can be taken as the supervision for optimizing PVP.

3.4. Loss Function

Three loss functions are employed to optimize the proposed method, including the visibility verification loss L_v for self-supervised visibility learning, the part-matching loss L_m for enhancing the relevance between corresponding parts, and the identity classification loss L_c for maintaining the discriminative power of each part feature. Therefore, the overall loss L can be formulated as,

$$L = L_v + L_c + L_m \tag{11}$$

Visibility Verification Loss L_v. We impose a Binary Cross Entropy loss for PVP module in training phrase with the self-supervision signal v^*, which is obtained via strategy mentioned in Sec. 3.3. Specifically, the product of part visibility scores of the input probe and gallery I^p, I^g are trained to approximate matching vector, which is formulated by:

$$L_v = -\sum_{i=1}^{N_p} v_i^* \log(\hat{v}_i^p \hat{v}_i^g) \tag{12}$$

where v_i^p and v_i^g correspond to the i-th part visibility score of probe and gallery images respectively.

Part Matching Loss L_m. After obtaining the optimal visibility score v^*, continue to optimize the matching quality function according to M enables to enhance the intra-part consistence. A part-based matching loss which is similar to the form of Eq.9 is employed here. By fixing v with value v^*, the matching loss is formulated as:

$$L_m = -v^{*T} M v^* + \lambda'^T v^* \tag{13}$$

Among this loss function, the first term could enhance the intra-part consistence and the second term enforces the network to extract complementary features from different parts, where $\lambda' \in \mathcal{R}^P$ is defined as,

$$\lambda_i' = \frac{\sum_{j=1, j\neq i}^{N_p} S_{ij}^p + S_{ij}^g}{2(N_p - 1)} \tag{14}$$

and S^p and S^g corresponding to the inter-part feature similarity matrix of probe and gallery, respectively.

Classification Loss L_c. To introduce discriminative power into the proposed network, we adopt a classification loss as the objective function. Following the construction of RPP [24], we fix the pre-trained PCB classifiers to maintain the knowledge learned under the uniform partition. Then the classification loss can be formulated as,

$$L_c = \sum_{i=1}^{N_p} CE(\hat{y}_i, y_i) \tag{15}$$

where CE is the Cross-Entropy loss, $\hat{y}_i$ is the prediction of the i-th part classifier, and y is the ground-truth ID.

4. Experiments

4.1. Datasets and Settings

Datasets. For experimental evaluation, we conduct experiments on two small-scale and one large-scale ReID benchmarks, including the Occluded-REID [36], the Partial-REID [31], and the large P-DukeMTMC-reID [36] dataset. Each reported occluded dataset is partitioned into two parts: the occluded person images and full-body person images. For model pre-training, we train the networks on the Market-1501 [30] dataset.

Algorithm 1 Pose-Guided Visible Part Matching.

```
Input: Training image data: I, $T$, $\lambda$.
Output: The parameters $\theta_e$, $\theta_v$, $\theta_a$ of PE , PVP and PGA
Initialize $\theta_e$,$\theta_v$, and $\theta_a$
for $t = 1, 2, \ldots, T$ do
    Randomly select a batch of images from I;
    Generate part feature $\{f_i\}$ with Eq. 3
    Predict visibility score $\{\hat{v}_i\}$ with Eq. 5
    Obtain pseudo label $\{v_i^*\}$ by solving Eq.9
    Update $\theta_v \leftarrow \frac{\partial}{\partial \theta_v} L_v$
    Update $\theta_a \leftarrow \frac{\partial}{\partial \theta_a} (L_c + L_m)$
    Update $\theta_e \leftarrow \frac{\partial}{\partial \theta_e} (L_v + L_c + L_m)$
end for
return $\theta_v$,$\theta_a$, and $\theta_e$
```

1) Occluded-REID [36] images are captured by mobile camera equipments in campus, including 2,000 annotated images belonging to 200 identities. Among the dataset, each person consists of 5 full-body person images and 5 occluded person images with various occlusions.

2) Partial-REID [31] includes 900 images of 60 pedestrians. Each person has 5 full-body person images, 5 occluded person images and 5 manually cropped partial person images from the occluded ones. In this work, we only use the full-body and occluded person images for evaluation.

3) P-DukeMTMC-reID [36] is a modified version based on DukeMTMC-reID dataset [32]. 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) Market-1501 [30] contains 32,668 labelled images of 1,501 identities observed from 6 cameras. The dataset is split into training set with 12,936 images of 751 identities and used for model pre-training only.

Evaluation Protocols. We report the Cumulated Matching Characteristics (CMC) [5] and mean Average Precision (mAP) [30] value for the proposed approach. The evaluation package is provided by [33], and all the experimental results are performed in a single query setting.

Implementation Details We take all occluded person images as probe set and full-body person images as gallery set on all three reported datasets. Specifically, for the Occluded-REID [36] and Partial-REID [31] datasets, due to the absence of the same prescribed split of training and test set, all the images are adopted for testing. With all training images resized as 384×128, we employed ResNet-50 [6] which is pre-trained with the same setting as PCB [24] to extract appearance features. This feature is then followed by a pose-guided attention pooling operation which generates N_p part features, where N_p is set as 6 by default. For pose estimation, we adopt the OpenPose [2] method pre-trained on the COCO dataset [13], which generates 18 keypoint heatmaps K and 38 part affinity fields L_p. The pro-

Table 2. Performance comparisons with the holistic and occluded methods on the three reported datasets. The 1st/2nd best results are in red and blue.

Method	Occluded-REID				Partial-REID				P-DukeMTMC-reID			
	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
IDE [30]	52.6	68.7	76.6	46.4	51.7	69.0	80.3	52.4	36.0	49.3	55.2	19.7
OsNet [34]	39.7	57.9	66.5	36.0	48.7	68.0	78.3	49.3	33.7	46.5	54.0	20.1
MLFN [4]	42.3	60.6	68.5	38.4	42.7	62.7	72.3	45.7	31.3	43.6	49.6	18.1
HACNN [10]	29.1	44.7	54.7	26.1	37.0	64.0	75.3	40.4	30.4	42.1	49.0	17.0
Part Bilinear [22]	54.9	70.8	77.7	50.3	57.7	77.3	85.7	59.3	39.2	50.6	56.4	25.4
PCB [24]	59.3	75.2	83.2	53.2	66.3	84.0	91.0	63.8	43.6	57.1	63.3	24.7
PCB+RPP [24]	55.8	74.4	81.2	51.3	63.7	82.3	90.0	61.2	40.4	54.6	61.1	23.4
Teacher-S [37]	55.0	64.5	77.3	59.8	69.2	76.6	85.8	73.1	18.8	24.2	32.2	22.4
PGFA [16]	57.1	77.9	84.0	56.2	68.0	82.0	86.7	56.2	44.2	56.7	63.0	23.1
PVPM	66.8	82.0	88.4	59.5	75.3	88.7	92.3	71.4	50.1	63.0	68.6	29.4
PVPM+Aug	70.4	84.1	89.8	61.2	78.3	89.7	93.7	72.3	51.5	64.4	69.6	29.2

posed PVP and PGA method is trained at a learning rate of 0.002 via the SGD optimizer. The training batch size, the training epoch, and the coefficient λ are set as 32, 30, and 0.9, respectively. This code is implemented under NVIDIA 1080Ti GPU environment and Pytorch platform.

4.2. Performance under Transfer Setting

Performance comparison under transfer setting is conducted by directly utilizing the model trained on Market1501 [30] without any further optimization.

Comparison with Holistic Methods The performance comparison with the holistic methods are illustrated in the first group of Table 2. Among these methods, HACNN [10] introduces the appearance-based attention mechanism into model training. Compared to the Part Bilinear [22] method which utilizes the pose information to improve the re-identification performance, the PCB(+RPP) [24] method propose to use a refined part pooling strategy. The '+Aug' corresponds to the result when training the PVPM model with images augmented with random occlusions to solve the data unbalance problem in the occluded training set. From the table, we can observe that the proposed method outperforms those holistic approaches by a large margin, with rank-1 accuracy surpasses the second-best holistic method by around 10% for all three reported benchmarks. This result may validate that, 1) it is essential to propose a specifically designed framework for the occluded ReID task; 2) matching with the visible parts shows better performance rather than using all parts.

Comparison with Occluded Methods We show the performance comparison with two specifically designed occluded ReID approaches in the second group of Table 2. The Teacher-S [37] proposes to train networks to learn a global feature with two auxiliary tasks, which would make networks paying more attention to person body parts. The PGFA [16] proposes a hard part matching method via a fixed region selection strategy and hand-crafted part visibility judgement method. Compared to these two approaches, our PVPM model achieves 70.4%, 78.3% and 51.5% at rank-1 on the Occluded ReID [36], Partial-REID [31] and P-DukeMTMC-reID [36] dataset, outperforming them by a large margin. This large performance improvement may be drawn from three aspects: 1) part matching works better for occluded ReID task rather than global feature learning; 2) a trainable part visibility prediction model could benefit more than the hand-crafted strategy; 3) training a high-level pose features can provide better guidance for person retrieval compared to simply fuse features with pose keypoint heatmaps;

Table 3. Comparison with partial ReID methods on the Partial-REID dataset. The 'manually crop' indicates the method use the original occluded images or the manually occlusion removed images for matching.

Method	rank-1	rank-3	manually crop
MTRC [12]	23.7	27.3	√
AWC+SWM [30]	37.3	46.0	√
SFR [7]	56.9	78.5	√
VPM [23]	67.7	81.9	√
PVPM	75.3	86.0	×
PVPM+Aug	78.3	87.7	×

Table 4. Performance on the P-DukeMTMC-reID dataset under supervised setting.

Method	Rank-1	Rank-5	Rank-10	mAP
Teacher-S [37]	51.4	50.9	-	-
IDE [30]	82.9	89.4	91.5	65.9
Baseline(PCB) [24]	79.4	87.1	90.0	63.9
PVPM	85.1	91.3	93.3	69.9

Comparison with Partial Methods Compared to occluded ReID task, partial ReID aims to solve the matching problem with the images manually cropped via a bounding box from the original images. This may result in image distortion and misalignment, and the occlusions still can not be totally removed, therefore, increasing the matching difficulties. In this section, four partial ReID methods are compared with the proposed PVPM in Table 3 on the Partial

Table 5. Performance comparisons with different component settings.

Method	Occluded-REID				Partial-REID				P-DukeMTMC-reID			
	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
Baseline(PCB) [24]	59.3	75.2	83.2	53.2	66.3	84.0	91.0	63.8	43.6	57.1	63.3	24.7
PVPM	66.8	82.0	88.4	59.5	75.3	88.7	92.3	71.4	50.1	63.0	68.6	29.4
PVPM-L_m	65.5	80.7	86.3	58.2	73.0	86.7	92.0	69.6	48.1	61.7	67.9	27.7
PVPM-thre	65.1	80.3	87.3	58.1	71.7	87.0	91.0	68.1	48.1	61.5	68.1	29.0
PGA only	61.1	77.2	84.5	55.0	68.7	85.0	91.7	65.6	43.9	58.1	64.5	26.6
PVP only	65.2	80.4	86.6	57.3	74.0	89.3	92.3	70.4	46.8	62.0	67.4	26.0
PVP only+Aug	69.0	83.5	88.4	60.9	74.7	89.0	92.3	71.2	49.7	63.3	69.3	27.5
PVPM+Aug	70.4	84.1	89.8	61.2	78.3	89.7	93.7	72.3	51.5	64.4	69.6	29.2

ReID dataset [31], listing the rank-1, rank-3 matching rates. We also demonstrate whether the model needs to match persons with the manually occlusion removed images or the original pictures. As can be seen, compared to those partial ReID methods, our PVPM+Aug model arrives 78.3% at rank-1, outperforming the second-best VPM [23] approach by 10.6%. Note that, our PVPM approach does not require to pre-process the images while testing, which shows better practicability in the real-world scenes.

4.3. Performance under Supervised Setting

For the large-scale dataset P-DukeMTMC-reID [36], we further run experiments to evaluate the performance when optimizing the model with the target training set. The results of two methods, IDE [30] as well as our part-based baseline method PCB [24] are demonstrated in Table 4. As can be observed, our PVPM method achieves 85.1% at rank-1, which surpasses the baseline method by 5.7%. This further illustrates our model superiority under the supervised setting for occluded person ReID.

4.4. Algorithm Analysis

In this subsection, we conduct experiments to thoroughly verify the effectiveness of the components of the Pose-Guided Attention (PGA) mechanism, the Pose-guided Visibility Prediction (PVP) model, the part matching loss L_m, the graph matching model and the augmented training samples with randomly generated occlusions. The experimental results on the reported three benchmarks are shown in Table 5. The 'Baseline' is the result of directly employing the PCB [24] model. The 'PGA only' means that we only use the pose-guided part features without further employment of part visibility computation. The 'PVP only' corresponds to the result that assigning each uniform part features with a visibility score without the soft pose-guided attention mask. The '-L_m' is the result when removing the part matching loss from the whole loss function. The '-thre' is the result when inferring pair visibility by thresholding their similarity. The '+Aug' indicates that we augment the training samples by randomly replacing a region in the image with a background patch, which is motivated by [36].

Table 6. Comparison results of generating part maps and visibility score from difference type of cues: appearance-based or pose-guided. PVPM is the proposed pose-guide method, 'RPP' indicates to refine the part maps from uniform spliting as in [24]. 'R+S' means the result when we further employ an appearance-based visibility predictor with the 'RPP'.

Datasets	Methods	Rank-1	Rank-5	Rank-10	mAP
Occluded	RPP	55.8	74.4	81.2	51.3
	R+S	51.8	69.3	76.6	47.3
	PVPM	66.8	82.0	88.4	59.5
Partial	RPP	63.7	82.3	90.0	61.2
	R+S	59.7	81.3	88.0	59.0
	PVPM	75.3	88.7	93.7	72.3
P-Duke	RPP	40.4	54.6	61.1	23.4
	R+S	35.6	47.9	53.3	21.1
	PVPM	50.1	63.0	68.6	29.4

From Table 5, we can observe that the employment of PGA block can achieve better performance. This suggests that the utilization of pose-guide attention do benefit the occluded re-identification task. When comparing the result of 'PVP only' and 'baseline', it can be easily drawn that computing a weighted distance according to the part visibility score improves the rank-1 performance by 5.9%, 7.7%, 3.2% on the three reported datasets. Note that, our graph model method outperforms the thresholding method, which demonstrates our model advantage as it considers the body part-to-part correlations while inferring their correspondence. What is more, when we remove the L_m from the entire loss function, performance drops by around 1-2% at rank-1 accuracy, validating its effectiveness. Besides, the result of the '+Aug' operation demonstrates that the augmented occluded training samples can make a contribution to performance gain.

4.5. Analysis of Pose Cues

Compared with the appearance cues, the pose-guided cues sometimes can provide more reliable information for occluded occasions. To validate the advantage of utilizing human pose information for part region generation and part visibility score prediction, we compare our PVPM model with an appearance-based part refine method RPP [24]. The quantitative result is illustrated in Table 6. As can be seen,

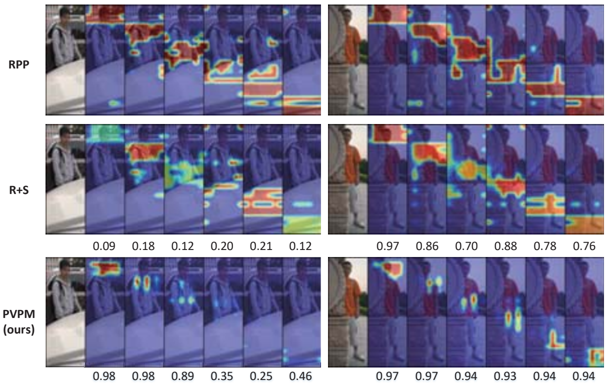

Figure 4. Visualization of part maps and visibility score generated from difference cues. The number under each picture denotes the predicted visibility score of this part. The pictures on each line indicate the part maps generated via our PVPM model, the appearance-based refined part model(RPP [24]), and the combination of RPP and a visibility score predictor(VSP), respectively

the appearance-based RPP [24] method does not achieve a performance boost on the occluded dataset comparing with the baseline method PCB [24] (in Table.5). Furthermore, we train a model with the same setting as our self-supervised framework but replace the PGA and PVP block with two appearance-based module, RPP [24] and a part visibility score predictor (VSP), respectively. This strategy is defined as 'R+S'. The further employment of the VSP method makes the performance drop further. For better viewing, we demonstrate the visualization result in Figure 4, including both the part maps and the predicted part visibility score. The visualization part maps show that the pose-guided attention mask can focus more on the regions which are not occluded. Therefore, we can deduce that, compared with pose cues, the appearance cues can not offer enough insight especially when facing new obstructions.

4.6. Parameter Analysis

The Impact of regularization coefficient λ. λ is the regularization coefficient of Eq.9. Small λ will weaken the discriminative ability of the visibility predictor, leading to all parts thought as visible. But a large λ may mislead the visibility predictor taking some local regions as unobservable. In this section, we compare the performance with different settings of λ, which varies from 0.6 to 1. We show the rank-1 accuracy and mAP variations in Figure.5. As can be seen, the performance reaches the peak value at 0.9, and drop a little bit with λ increasing to 1. This performance trend just validates our expectation of the coefficient λ.

The Impact of Part Number N_p. N_p determines the granularity of the part feature. We conduct several experiments by setting N_p from 2 to 8, and demonstrate the result in Figure 6, including the rank-1 matching rate and the mAP value. As can be seen, with N_p increases, the performance keeps improves at first, and reaches the peak when N_p arrives 4. However, the performance starts to drop with the part number continuing enlarging to 8. We suggest that this phenomenon may be drawn by the over-increased N_p, making the small parts becoming similar to each other and decreasing the discriminative ability of our model.

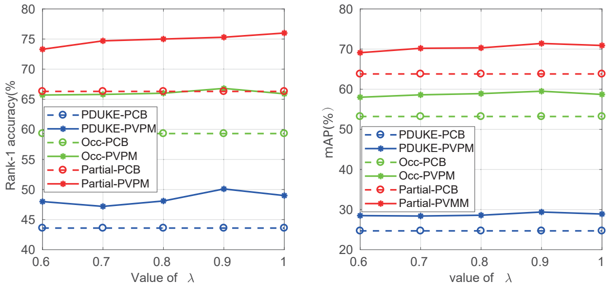

Figure 5. Rank-1 accuracy and mAP with different λ settings. The red, green and blue lines correspond to the results of the Partial ReID, Occluded ReID and p-DukeMTMC-reID dataset.

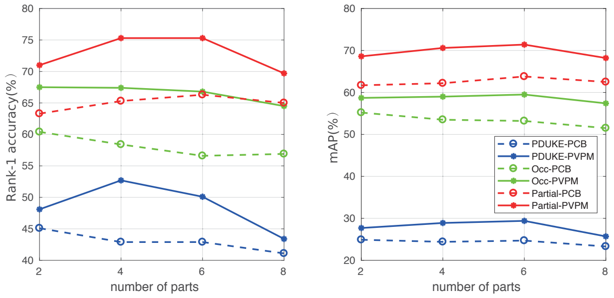

Figure 6. Rank-1 accuracy and mAP comparison with different setting of part number N_p.

5. Conclusion

In this paper, we propose a novel Pose-guided Visible Part Matching (PVPM) algorithm for occlusion ReID task. The proposed PVPM jointly considers the discriminative pose-guided attention and part visibility in a unified framework. Unlike most existing methods which utilize visibility cues from other data source directly, we explore the part correspondence on target data and self-mine visibility score via graph matching. A self-learning method was introduced for pseudo label generation and optimize visibility predictor without data bias. Sufficient experimental results on the three reported occluded datasets demonstrate the superiority of the proposed model for occluded person ReID task.

Acknowledgements. The paper is supported in part by the National Key R&D Program of China under Grant No.2018AAA0102001 and National Natural Science Foundation of China under Grant No.61725202, U1903215, 61829102, 91538201, 61751212 and the Fundamental Research Funds for the Central Universities under Grant No. DUT19GJ201.

References

[1] Honglong Cai, Zhiguan Wang, and Jinxing Cheng. Multi-scale body-part mask guided attention for person re-identification. In *CVPR Workshops*, 2019.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

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

[5] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007.

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

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

[8] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction, 2018.

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

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

[11] Wei-Hong Li, Fa-Ting Hong, and Wei-Shi Zheng. Learning to learn relation for important people detection in still images. In *CVPR*, 2019.

[12] Shengcai Liao, Anil K Jain, and Stan Z Li. Partial face recognition: Alignment-free approach. *TPAMI*, 2012.

[13] 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.

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

[15] 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.

[16] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. 2019.

[17] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[18] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *CVPR*, 2018.

[19] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *CVPR*, 2018.

[20] 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.

[21] Yumin Suh, Kamil Adamczewski, and Kyoung Mu Lee. Subgraph matching using compactness prior for robust feature correspondence. In *CVPR*, 2015.

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

[23] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, 2019.

[24] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *ECCV*, 2018.

[25] Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *CVPR*, 2018.

[26] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017.

[27] 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.

[28] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

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

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

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

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

[33] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv preprint arXiv:1910.10093*, 2019.

[34] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. *arXiv preprint arXiv:1905.00953*, 2019.

[35] Qin Zhou, Heng Fan, Shibao Zheng, Hang Su, Xinzhe Li, Shuang Wu, and Haibin Ling. Graph correspondence transfer for person re-identification. In *AAAI*, 2018.

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

[37] Jiaxuan Zhuo, Jianhuang Lai, and Peijia Chen. A novel teacher-student learning framework for occluded person re-identification. *arXiv preprint arXiv:1907.03253*, 2019.