AONet: Attentional Occlusion-aware Network for Occluded Person Re-identification

Guangyu Gao$^{1[0000-0002-0083-3016]}$, Qianxiang Wang$^{1[0000-0002-1322-2476]}$,
Jing Ge$^{1[0000-0001-9114-690X]}$, and Yan Zhang$^{1[0000-0002-9170-7743]}$

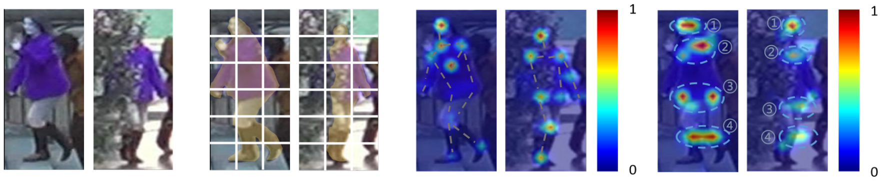
Beijing Institute of Technology, Beijing, China
guangyugao@bit.edu.cn

Abstract. Occluded person Re-identification (Occluded ReID) aims to verify the identity of a pedestrian with occlusion across non-overlapping cameras. Previous works for this task often rely on external tasks, *e.g.*, pose estimation, or semantic segmentation, to extract local features over fixed given regions. However, these external models may perform poorly on Occluded ReID, since they are still open problems with no reliable performance guarantee and are not oriented towards ReID tasks to provide discriminative local features. In this paper, we propose an Attentional Occlusion-aware Network (AONet) for Occluded ReID that does not rely on any external tasks. AONet adaptively learns discriminative local features over latent landmark regions by the trainable pattern vectors, and softly weights the summation of landmark-wise similarities based on the occlusion awareness. Also, as there are no ground truth occlusion annotations, we measure the occlusion of landmarks by the awareness scores, when referring to a memorized dictionary storing average landmark features. These awareness scores are then used as a soft weight for training and inferring. Meanwhile, the memorized dictionary is momenta updated according to the landmark features and the awareness scores of each input image. The AONet achieves 53.1% mAP and 66.5% Rank1 on the Occluded-DukeMTMC, significantly outperforming state-of-the-arts without any bells and whistles, and also shows obvious improvements on the holistic datasets Market-1501 and DukeMTMC-reID, as well as the partial datasets Partial-REID and Partial-iLIDS. The code and pre-trained models will be released online soon.

Keywords: Occluded ReID · Occlusion-aware · Landmark, Orthogonal.

1 Introduction

Most Person Re-identification (ReID) [12, 2, 31] approaches focus more on holistic pedestrian images, and tend to fail in real-world scenarios where a pedestrian is partially visible, *e.g.*, occluded by other objects. The Occluded person Re-identification (Occluded ReID) is then investigated which aims to handle the occlusion distractions. Some previous Occluded ReID methods perform part-to-part matching based on fine-grained external local features [13, 25], *e.g.*, with body parts assigned larger weights and occlusion parts smaller weights.



(a) Holistic and Occluded Images (b) Part Features by Segmentation (c) Body Parts by Pose Estimation (d) Landmarks by the OANet

Fig. 1. Illustration of local feature responses. (a) Holistic and occluded images. (b) Local features by partitioning over segmentation mask. (c) Local features based on pose estimation. (d) Landmark features by our AONet.

The key to solving Occluded ReID is to locate landmark regions and then extract well-aligned features from non-occluded landmark regions, while reasonably reducing or prohibiting the use of features from occluded landmark regions. Some Occluded ReID works use body parts attained from pose estimation for local feature extraction [25, 3, 4], and suppress or exclude the local features of some occluded body parts with low pose confidence. However, the reliability of pose estimation is not guaranteed (*e.g.*, failure on the knees and waist in Fig. 1 (c)). Moreover, pose features are often not necessarily adapted to ReID tasks due to cross-task variance. Another group of methods [23, 13, 21, 33] extracts local features directly on uniformly partitioned grids on pedestrian images, and measures the occlusion of each grid guided by the semantic segmentation task, as shown in Fig. 1 (b). However, due to the different poses and non-rigid deformation of the human body, these models cannot accurately perform part alignment and thus often fail. In addition, there are some methods that achieve Occluded ReID by locating occluded parts or measuring the occlusion degree using the pose estimation [28, 4] or semantic segmentation tasks [33, 3].

However, the aforementioned methods rely on external tasks, such as pose estimation or semantic segmentation, to extract local features on fixed given regions of the human body. On one hand, the results of these external tasks may be imprecise; on the other hand, the obtained local features are usually not discriminative enough for Occluded ReID. [10] presented the Matching on Sets (MoS), positioning Occluded ReID as a set matching task without using external models. Compared to this work, we go further to adaptively extract more discriminative local features as well as more accurately sense and measure the occlusions. We then propose an Attentional Occlusion-aware Network (AONet) with the Landmark Activation Layer and the Occlusion Awareness (OA) component. The latent landmark features refer to features of ReID oriented local parts (*i.e.*, latent landmarks), and are resistant to landmark occlusion. The occlusion awareness score measures the visibility of each landmark according to the average landmark features in the memorized dictionary. Besides, to prevent the model collapse problem that multiple landmarks focus on the same region, we involve the orthogonality constraints among landmarks features.

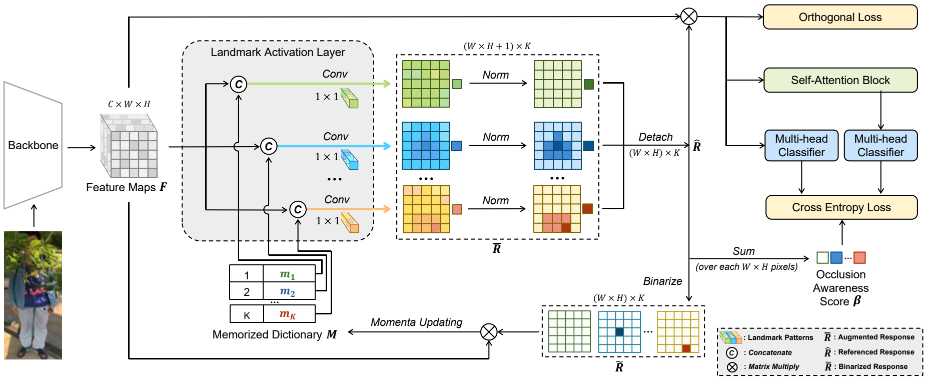


Fig. 2. The framework of AONet, including a Landmark Activation (LA) layer to extract the landmark features, and the Occlusion Awareness (OA) score to measure the occlusion. The responses of occluded pixels will be lower than the corresponding average response passing through the LA layer. Then, the normalization over all pixels and the average responses will further scale down these occluded responses (the green branch). Finally, the normalized pixel responses are summed up as the occlusion awareness score, and used to update the memorized dictionary.

Our main contributions can be summarized as follows:

- Instead of relying on any external tasks, we only use a learnable parameter matrix (*i.e.*, the landmark patterns) and a memorized dictionary storing the average landmark features, to guide the extraction of landmark features that are more discriminative and resistant to occlusion.
- Furthermore, we define the occlusion awareness score to sense and measure the occlusion of each landmark explicitly, especially by referring to the average landmark features in the memorized dictionary.
- Our AONet achieves excellent performance on not only the occluded dataset Occluded-DukeMTMC, but also the holistic and partial datasets, *i.e.*, DukeMTMC-reID, Market-1501, Partial-REID, and Partial iLIDS, significantly outperforming state-of-the-art.

2 Related Works

Person ReID has been studied in terms of both feature representation learning [34, 23, 30] and distance metric learning [24, 1, 8]. However, most ReID methods focus on matching the holistic pedestrian images, and do not perform well on occlusion images [13, 25], which limits their applicability in real-world scenarios.

Occluded ReID [25, 10] is aimed at matching occluded person images to holistic ones across dis-joint cameras, which is challenging due to distracting factors like cluttered scenes or dense crowd. To solve it, [7] proposed an occlusion-robust alignment-free model, using an occlusion-sensitive foreground probability generator with guidance from a semantic segmentation model. [13] refined the setup of the Occluded ReID problem to be more realistic, *i.e.*, both probe and gallery images containing occlusion. They introduced a PGFA method that exploits pose landmarks to disentangle useful information from the occlusion noise. Here, we tackle an Occluded ReID problem as defined in [13].

Later, [4] proposed a PVPM method to jointly learn the pose-guided local features and the self-mined part visibility. [25] proposed an HOReID method to learn high-order relations for discriminative features and topology information for robust alignment, by an external human key-points prediction model. In [33], a Semantic-aware Occlusion-Robust Network (SORN) was proposed that exploits the intrinsic relationship between person ReID and semantic segmentation. Also, [14] proposed a Semantic Guided Shared Feature Alignment (SGSFA) method to extract features focusing on the non-occluded parts, using guidance from external human parsing and pose estimation models. The above works require guidance information from external tasks (*e.g.*, semantic segmentation, pose estimation) either for local feature extraction or occlusion measurement. Recently, [10] presented the Matching on Sets (MoS) method, viewing Occluded ReID as a set matching task without requiring spatial alignment.

3 Attentional Occlusion-aware Network

The Attentional Occlusion-aware Network (AONet) mainly includes the extraction of the attentional landmark features, and the calculation of the Occlusion Awareness (OA) score, as shown in Fig. 2. Meanwhile, a learnable matrix is used to explicitly represent the *landmark patterns* for the more discriminative features. A *memorized dictionary* is defined as a strong reference, which stores the average landmark features and is dynamically updated in a momentum way. The discriminative local features, *i.e.*, the landmark features, are extracted adaptively according to both the memorized dictionary and the landmark patterns.

3.1 Landmark Patterns & Memorized Dictionary

Landmark Patterns. We define the *landmark patterns* $\boldsymbol{I} \in \mathbb{R}^{C\times K}$ as trainable parameters to attend to specific discriminative landmarks, *i.e.*, the attentional latent landmarks. We expect the learned *landmark patterns* to encode local patterns, which help explain the inputs (feature maps $\boldsymbol{F}$).

Memorized Dictionary. We also define the *memorized dictionary* $\boldsymbol{M} \in \mathbb{R}^{C\times K}$ to store the average features of the K latent landmarks. $\boldsymbol{M}$ is zero-initialized but momentum updated under the guidance of landmark patterns batch by batch. Moreover, the updating considers the occlusion of each landmark, *i.e.*, using the referenced response maps in the calculation of the occlusion awareness scores (see

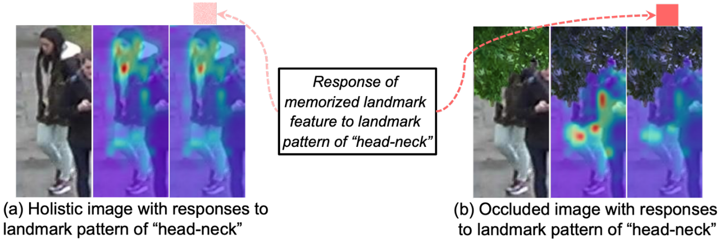

Response of memorized landmark feature to landmark pattern of "head-neck"
(a) Holistic image with responses to landmark pattern of "head-neck"
(b) Occluded image with responses to landmark pattern of "head-neck"

Fig. 3. Visualization of the effect *with* or *without* the reference of *average response* on the example pattern 'head-neck'. The use of *average response* has no particular impact on the holistic image in (a), but is effective in suppressing false alarms on the occluded image in (b). In (b), facing the occlusion of 'head-neck', the response map has false alarms on 'knees' without referring to average response (the 2nd image), but this gets obviously alleviated with the reference (the 3rd image).

details in Sec. 3.4). Namely, given the *referenced response map* $\widehat{\boldsymbol{R}}_k \in \mathbb{R}^{W\times H}$ for the k_{th} landmark, we binarize $\widehat{\boldsymbol{R}}_k$ as $\tilde{\boldsymbol{R}}_k \in \mathbb{R}^{W\times H}$ by setting all pixels corresponding to the maximum value to 1 and the rest to 0. Then, given a batch of B images, we use momentum updating to get the updated memorized dictionary $\boldsymbol{M}_k^{t+1}$ with a balance weight (α) as:

$$\boldsymbol{M}_k^{t+1} = \alpha \boldsymbol{M}_k^t + (1-\alpha)\frac{1}{B}\sum\nolimits_{b=1}^{B} \boldsymbol{F}_b \tilde{\boldsymbol{R}}_k. \tag{1}$$

3.2 Attentional Latent Landmarks

The learnable landmark patterns $\boldsymbol{I} = \{i_k\}_{k=1}^K, i_k \in \mathbb{R}^C$ should be trained together with other parameters of the network. The 1x1 convolution can be seen as an operation where a $1\times 1\times C$ sized filter is applied over the input and then weighted to generate several activation maps. That is, the 1×1 filter can be thought of as some type of pattern matching to create a linear projection of a stack of feature maps. Therefore, we realize the landmark patterns by the 1×1 filters, as shown in the Landmark Activation layer of Fig. 2.

In details, the 1x1 convolution layer appended after the CNN backbone network takes $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ (feature maps of an input image) as input, and outputs K landmark-specific response maps $\boldsymbol{R} = \{\boldsymbol{R}_k\} \in \mathbb{R}^{K\times W\times H}$. We normalize these response maps among all pixels to form the basic normalized response maps $\check{\boldsymbol{R}} \in \mathbb{R}^{K\times W\times H}$, then the value of pixel (w,h) in the k_{th} map is calculated as,

$$\check{\boldsymbol{R}}_k(w,h) = \frac{\phi(i_k, \boldsymbol{F}(w,h))}{\sum_{(i,j)=(1,1)}^{(W,H)} \phi(i_k, \boldsymbol{F}(i,j))}, \tag{2}$$

and $\phi(i_k, \boldsymbol{F}(i,j)) = \exp(i_k^T \boldsymbol{F}(i,j))$ is the similarity based response.

After that, without considering occlusion awareness, we easily obtain the Standard Landmark-specific (SL) features of $\bar{\boldsymbol{f}}_k \in \mathbb{R}^C$ for the k_{th} landmark by

Algorithm 1 The Main Flowchart of the AONet.

```
Input: Batch of feature maps $\boldsymbol{\mathcal{F}} = \{\boldsymbol{F}^b \in \mathbb{R}^{W\times H\times C}\}_{b=1}^{B}$; landmark patterns $\boldsymbol{I} = \{i_k\} \in \mathbb{R}^{C\times K}$; where $B$, $C$ and $K$ are the size of batch, channel and the number of landmarks.
Output: The awareness scores $\boldsymbol{\beta} = \{\beta_k\}$ and updated memorized dictionary $\boldsymbol{M}$;
1: Initialize the memorized dictionary $\boldsymbol{M}_t = \{m_k\} \in \mathbb{R}^{C\times K}$;
   (noting: superscript $b$ is omitted until step 11 for convenience.)
2: for $t = 1$ to $T$ do
3:     Response maps $\boldsymbol{R}$=$\phi(\boldsymbol{I}, \boldsymbol{F}) \in \mathbb{R}^{K\times W\times H}$, where $\boldsymbol{R}_k \in \mathbb{R}^{W\times H}$ is the $k_{th}$ landmark's response on $\boldsymbol{F}$;
4:     Average responses $\boldsymbol{a} = \{a_k\}$, where $a_k = \phi(i_k, m_k)$.
5:     Each augmented response map $\overline{\boldsymbol{R}}_k = \{\boldsymbol{R}_k; a_k\} \in \mathbb{R}^{W\times H+1}$.
6:     Normalizing $\overline{\boldsymbol{R}}_k$ with Eqn. 3.
7:     Referenced response map $\widehat{\boldsymbol{R}}_k = \{\widehat{\boldsymbol{R}}_k(w,h)\} \in \mathbb{R}^{W\times H}$, i.e., detaching the value corresponding to $s_k$.
8:     Calculating the awareness scores (e.g., $\beta_k^b$) based on Eqn. 4;
9:     The $k_{th}$ OA feature $\boldsymbol{f}_k = \boldsymbol{F}\widehat{\boldsymbol{R}}_k$.
10:    Binarization over $\widehat{\boldsymbol{R}}_k$ to get binarized response map $\tilde{\boldsymbol{R}}_k$.
11:    Updating the $K$ memorized landmark features:
       $\boldsymbol{M}_k^{t+1} = \alpha\boldsymbol{M}_k^t + (1-\alpha)\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{F}_b\tilde{\boldsymbol{R}}$;
12: end for
13: return $\boldsymbol{\beta}$ and $\boldsymbol{M}$ respectively;
```

$\bar{\boldsymbol{f}}_k = \boldsymbol{F}\check{\boldsymbol{R}}_k$. However, the SL features cannot accurately reflect the response of landmarks in the occluded image. As shown in Fig. 3 (b), when the example landmark (seems to be the 'head-neck' parts) is occluded, this landmark still has large activated regions (*i.e.*, the false alarm on the parts of 'knees'). Thus, we adopt the landmark features that characterize the occlusion awareness, *i.e.*, the OA features (see Sec. 3.4) instead of the SL features finally.

3.3 Referenced Response Map

Meanwhile, a special feature map, *i.e.*, the *referenced response map*, is defined to measure the occlusion awareness and represent the discriminative feature. We first calculate the similarity-based response between each landmark pattern (*e.g.*, $i_k \in \boldsymbol{I}$) and its corresponding memorized average feature (*e.g.*, $m_k \in \boldsymbol{M}$), which is named as the *average response* (*e.g.*, $a_k = \phi(x_k, m_k)$). While the memorized average features are the statistical representation of each landmark, the *average responses* can be used as some real and strong reference values to suppress false alarms, *e.g.*, scaling down responses of false alarms through uniform normalization, as shown in Fig. 3. More details can be seen in Alg. 1.

Then, for an input image, given a landmark-specific response map $\boldsymbol{R}_k = \phi(i_k, \boldsymbol{F})$, if all responses in $\boldsymbol{R}_k$ are significantly lower than the average response a_k, it means that this landmark is not present in this image, and the area corresponding to this landmark is occluded. Thus, we concatenate each a_k to the corresponding $\boldsymbol{R}_k$ to form the augmented response map $\overline{\boldsymbol{R}}_k$, which is then normalized in a similar way as Eqn. (2) (but on $W \times H + 1$ elements). That is, the normalized response of the k_{th} landmark pattern on the n_{th} pixel by the similarity function ϕ is calculated by

$$\widehat{\boldsymbol{R}}_k(w,h) = \frac{\phi(i_k, \boldsymbol{F}(w,h))}{a_k + \sum_{(i,j)=(1,1)}^{(W,H)} \phi(i_k, \boldsymbol{F}(i,j))}. \tag{3}$$

Given a landmark pattern i_k, pixel responses $\phi(i_k, x_n)$ that are far below the average response a_k are normalized to small values or even 0. Consequently, the responses of the occluded pixels, and pixels unrelated to the landmark patterns i_k are greatly suppressed, as shown in Fig. 3. Finally, we detach out the normalized response value of the average features, and rename the remained part as the *referenced response map*, *i.e.*, $\widehat{\boldsymbol{R}}_k = \{\widehat{\boldsymbol{R}}_k(w,h)\} \in \mathbb{R}^{W \times H}$.

3.4 Occlusion Awareness

Although we do not have true annotations about the occlusion of each landmark, we can utilize the memorized dictionary $\boldsymbol{M}$ that stores the average feature of each landmark, as a special strong reference to measure the occlusion, as shown in Alg. 1. Namely, the pixels that refer to a particular landmark should have a large response to this landmark's pattern. Namely, the feature of pixels referring to a particular landmark should be similar to the memorized average feature of that landmark, and also, both of them have comparable similarity based responses to the corresponding landmark pattern.

Occlusion Awareness Score. Ideally, if the regions referring to a landmark are occluded, there should be no responses to this landmark, and all pixels should not be used for learning any learnable landmark pattern. However, the network itself is not aware of the occlusion, and the responses of this landmark (*e.g.*, 'head-neck') will transfer to other unoccluded but wrong regions (*e.g.*, 'knees') to extract features (see Fig. 3). An intuitive idea for addressing this problem is to accurately measure the degree of occlusion by some metric (*i.e.*, awareness score), and then use it to suppress the impact of occluded regions in training and inference. Therefore, we further explicitly define an occlusion awareness score to measure the degree of occlusion based on the *referenced response map* $\widehat{\boldsymbol{R}}$. Specifically, we define the awareness score of the k_{th} landmark as,

$$\beta_k = \sum_{(1,1)}^{(W,H)} \widehat{\boldsymbol{R}}_k(w,h). \tag{4}$$

Then, β_k is used to reduce the weight of the occluded landmarks not only in training but also in inference.

Occlusion Awareness Feature. We need not only to sense the awareness score of each landmark but also involve such occlusion awareness in feature representation. Thus, the *referenced response maps* are also used as very crucial guidance to generate the more discriminative landmark features, *i.e.*, Occlusion Awareness (OA) features. Specifically, we replace $\check{\boldsymbol{R}}_k$ to the referenced response map $\widehat{\boldsymbol{R}}_k$, and get the k_{th} OA feature as $\boldsymbol{f}_k = \boldsymbol{F}\widehat{\boldsymbol{R}}_k$.

3.5 Training and Inference

Training Losses. We use the cross-entropy loss weighted by the occlusion awareness to constrain each landmark. Specifically, we perform the classification loss on both the OA features before and after the self-attention block, *i.e.*,

$$L_{cls} = -\frac{1}{K}(\sum_{k=1}^{K}\beta_k \log p_k^1 + \sum_{k=1}^{K}\beta_k \log p_k^2), \tag{5}$$

where β_k is the k_{th} landmark's awareness score, and p_k^1 and p_k^2 are the predicted probability of the k_{th} landmark features before and after the self-attention block.

Without any other constraints but only the classification loss, different landmarks are easy to collapse to focus on the same part. Thus, we propose the *orthogonal loss* to ensure spatial diversity between landmark features. In detail, when the cosine similarity is calculated between two landmark features ($\boldsymbol{f}_i$ and $\boldsymbol{f}_j$), the *orthogonal loss* is defined as

$$L_{ot} = -\frac{1}{K^2}\sum_{i=1}^{k}\sum_{j=i+1}^{K}\log(1 - |\mathrm{cosine}(\boldsymbol{f}_i, \boldsymbol{f}_j)|_+). \tag{6}$$

where $|\cdot|_+$ means the ramp function, *i.e.*, $max(0, \cdot)$.

Finally, the overall objective function is formulated by

$$L_{AONet} = L_{cls} + \lambda_{ot}L_{ot}, \tag{7}$$

where L_{cls} is the cross-entropy based classification loss, and L_{ot} refers to the *orthogonal loss* among landmark features before the self-attention block, and λ_{ot} is the balance weight.

Inference. For inference, given a pair of images ($\boldsymbol{im}^1$ and $\boldsymbol{im}^2$) and their feature maps ($\boldsymbol{F}^1$ and $\boldsymbol{F}^2$), as well as their landmark features (*e.g.*, $\boldsymbol{f}_k^1$ and $\boldsymbol{f}_k^2$), their similarity is calculated based on the cosine similarity $cosine(\cdot)$ by

$$\mathrm{sim}(\boldsymbol{im}^1, \boldsymbol{im}^2) = \frac{1}{K}\sum\nolimits_{k=1}^{K}\beta_k^1\beta_k^2 cosine(\boldsymbol{f}_k^1, \boldsymbol{f}_k^2), \tag{8}$$

where β_k^1 and β_k^2 are the occlusion awareness scores of the k_{th} landmark.

4 Experiments

4.1 Datasets and Implementations

We mainly evaluate AONet on the most popular occluded ReID dataset, *i.e.*, Occluded-DukeMTMC [13], where both the probe and gallery images have occlusion. In addition, we also experiment on holistic person ReID datasets: Market-1501 [35] and DukeMTMC-reID [15], as well as the partial ReID datasets:

Table 1. Comparison of performance on metrics of Ranks and mAP on the Occluded-DukeMTMC dataset.

Methods	Rank1	Rank5	Rank10	mAP
PGFA$_{[\text{ICCV19}]}$	51.4	68.6	74.9	37.3
HOReID$_{[\text{cvpr20}]}$	55.1	-	-	43.8
SORN$_{[\text{TCSVT20}]}$	57.6	73.7	79.0	46.3
SGSFA$_{[\text{ACML20}]}$	62.3	77.3	82.7	47.4
DIM$_{[\text{arXiv17}]}$	21.5	36.1	42.8	14.4
PartAligned$_{[\text{ICCV17}]}$	28.8	44.6	51.0	20.2
RandErasing$_{[\text{AAAI20}]}$	40.5	59.6	66.8	30.0
HACNN$_{[\text{CVPR18}]}$	34.4	51.9	59.4	26.0
AOS$_{[\text{CVPR18}]}$	44.5	-	-	32.2
PCB$_{[\text{ECCV18}]}$	42.6	57.1	62.9	33.7
PartBilinear$_{[\text{ECCV18}]}$	36.9	-	-	-
FD-GAN$_{[\text{NeurIPS18}]}$	40.8	-	-	-
DSR$_{[\text{CVPR18}]}$	40.8	58.2	65.2	30.4
MoS$_{[\text{AAAI21}]}$	61.0	-	-	49.2
AONet	**66.5**	**79.4**	**83.8**	**53.1**
MoS$_{w/ibn\,[\text{AAAI21}]}$	66.6	-	-	55.1
AONet$_{w/ibn}$	**68.8**	**81.4**	**85.8**	**57.3**

Partial-REID [36] and Partial-iLIDS [6]. All experiments are performed based on a single query image and without re-ranking [37].

We use ResNet50 [5] pre-trained on ImageNet as the backbone network. For a fair comparison, we also incorporate the *instance batch normalization* (*ibn*) into ResNet50 (*i.e.*, AONet$_{w/ibn}$) as in [10]. To acquire high-resolution feature maps, the stride of conv4_1 is set to 1. We resize original images into 256×128, with a half probability of flipping them horizontally. Then, the images are padded by 10 pixels and randomly cropped back to 256×128, and then randomly erased with a half probability. We use the Adam optimizer [11] with a learning rate of $3.5e-4$, warm up the training in the first 20 epochs and decay the learning rate with 0.1 in the $50th$ and $90th$ epoch. The batch size is 64, 4 images per person, and a total of 120 epochs are trained end-to-end. The weight of orthogonal loss, *i.e.*, λ_{ot}, is set to 0.01 and the momentum α for memorized dictionary updating is set to 0.9. If not specified, the number of landmarks is set as 6.

4.2 Comparisons to State-of-the-arts

Results on Occluded ReID Dataset. Table 1 shows the performance of AONet and several competing methods, including *methods without external models*: DIM [32], PartAligned [34], RandErasing [38], HACNN [12], AOS [9], PCB [23], PartBilinear [19], FD-GAN [18], DSR [6], MoS [10], *methods with external models*: PGFA [13], SORN [33], SGSFA [14], HOReID [25], and the most related *set matching based method* [10], on the Occluded-DukeMTMC dataset.

Table 2. Performance comparison on Holistic Person ReID datasets of the Market-1501 and DukeMTMC-reID.

Methods	Market-1501		DukeMTMC-reID	
	Rank1	mAP	Rank1	mAP
PCB+RPP[ECCV18]	93.8	81.6	83.3	69.2
MGN[MM18]	95.7	86.9	88.7	78.4
VPM[CVPR19]	93.0	80.8	83.6	72.6
SORN [TCSVT20]	94.8	84.5	86.9	74.1
PDC[ICCV17]	84.2	63.4	-	-
PSE[CVPR18]	87.7	69.0	27.3	30.2
PGFA[ICCV19]	91.2	76.8	82.6	65.5
HOReID[CVPR20]	94.2	84.9	86.9	75.6
PartAligned[ICCV17]	81.0	63.4	-	-
HACNN[CVPR18]	91.2	75.6	80.5	63.8
CAMA[CVPR19]	94.7	84.5	85.8	72.9
MoS[AAAI21]	94.7	86.8	88.7	77.0
AONet	95.2	86.6	88.7	77.4

Our AONet shows a significant advantage over other methods. Note that our AONet uses no external models as in most of the previous works. Moreover, on Rank1 and mAP, AONet improves 4.2% and 5.7% over the SOTA method *SGSFA* (with external models). AONet also improves 5.5% and 3.9% over *MoS* (without external models). In MoS, a pre-trained backbone network IBN is used to achieve better results, so we also propose the $\mathbf{AONet}_{w/ibn}$ utilizing IBN, which also achieves better results.

Results on Holistic ReID Datasets. Many related methods achieved good performance on Occluded ReID datasets, but they perform unsatisfactorily on holistic person ReID datasets and cannot be applied widely [13]. The AONet is also evaluated on the holistic person ReID datasets (*i.e.*, Market-1501 and DukeMTMC-reID) and compared with three groups of competing methods: *uniform-partition based* (PCB [23], VPM [22], MGN [26], *pose-guided based* (PDC [17], PSE [16], PGFA [13], HOReID [25]) and *attention-guided based* methods (PartAligned [34], HACNN [12], CAMA [29]). As shown in Table 2, the AONet produces satisfactory results in holistic cases even using an occluded oriented network. Meanwhile, the methods of different groups all perform well on holistic datasets and without large performance gaps. The reason could be, that almost all body parts are visible in holistic datasets, offering a greater possibility to locate all parts and thus obtain discriminative features easily. Meanwhile, AONet not only achieves SOTA performance on the occluded dataset, but also achieves competitive results on holistic datasets.

Results on Patial ReID Datasets. To fully validate the effectiveness of the AONet, we also conduct experiments on the partial person ReID datasets of Partial-REID and Partial-iLIDS. Since these two datasets are always only used as

Table 3. Performance comparison on Partial ReID datasets of Partial-REID and Partial-iLIDS.

Methods	Partial-REID		Partial-iLIDS	
	Rank1	Rank3	Rank1	Rank3
DSR[CVPR18]	50.7	70.0	58.8	67.2
VPM[CVPR19]	67.7	81.9	65.5	84.7
HOReID[CVPR20]	**85.3**	91.0	72.6	86.4
AONet (*crop*)	85.0	**92.7**	68.1	84.9
PGFA[ICCV19]	68.0	80.0	69.1	80.9
SGSFA[ACML20]	68.2	-	-	-
SORN [TCSVT20]	76.7	84.3	79.8	86.6
AONet (*whole*)	75.3	86.3	**80.7**	**86.6**

the test dataset, existing works are trained on other ReID dataset (*e.g.*, Market-1501). Not only that, but existing works also use the partial ReID dataset in two different ways (the two groups in Table 3), the main difference being whether the visible pedestrian area is cropped out separately as a new image. For example, the SOTA methods of HOReID [25] and SORN [33] are evaluated on partial datasets with images of the whole pedestrian or the cropped visible parts, respectively. We use **AONet**(*whole*) and **AONet**(*crop*) to refer to the performance of AONet on partial datasets in these two ways. As shown in Table 3, our **AONet**(*whole*) and **AONet**(*crop*) achieve the best performance on datasets of Patial-REID and Partial-iLIDS respectively, which proved the efficiency of our approach.

4.3 Ablation Studies

We test the effect of components in AONet with the below variants on Occluded-DukeMTMC: i) *SLFea* (Standard Landmark-specific features), extracting landmark features without occlusion awareness (Eqn. (**??**)). ii) *SAtt* (Self-Attention), enabling information interaction between landmark features. iii) *OAFea* (Occlusion Awareness Features), being similar to *SLFea*, but referring to the memorized features (Eqn. (**??**)). iv) *OAScore* (Occlusion Awareness Score), measuring the occlusion degree of each landmark. v) *OLoss* (orthogonal loss), constraining over different pairs of landmark features.

With or without each component. We define a basic baseline *Base_ GAP*, which includes the backbone of ResNet, a Global Average Pooling (GAP) layer, and a softmax layer. As shown in Table 4, compared to *Base_ GAP*, utilizing *SLFea* achieves significantly better performance, reflecting the advantage of set matching over global feature matching. Utilizing *SAtt* gains an extra improvement of 1.3% on Rank1 and 1.4% on mAP. Besides, with the simple combination of *SLFea* and *SAtt*, Rank1 of 58.6% is achieved, better than most previous methods as shown in Table 1. We argue that our attentional landmarks would facilitate reconstructing the information of the occluded landmark using other land-

Table 4. The ablation study of the components in AONet.

Method	SLFea	SAtt	OAFea	OAScore	OLoss	Rank1	mAP
Base_GAP						49.4	40.0
Base_SLFea	✓					57.3	43.9
Base_SAtt	✓	✓				58.6	45.3
AONet†		✓	✓			62.6	51.3
AONet‡		✓	✓	✓		65.6	52.5
AONet		✓	✓	✓	✓	**66.5**	**53.1**

Table 5. Ablation study of the components compared to baselines on the Occluded-DukeMTMC dataset.

Methods	Rank1	Rank5	Rank10	mAP
Base_GAP	49.4	63.7	68.9	40.0
Base_Pose	52.1	66.2	71.1	42.3
Base_OAFea (AONet†)	62.6	77.1	81.6	51.3
AONet†	62.6	77.1	81.6	51.3
+Base_Max	63.4	77.3	82.1	51.7
+OAScore (AONet‡)	65.6	79.1	83.6	52.5
AONet‡	65.6	79.1	83.6	52.5
+Base_RegL	65.9	78.8	83.7	52.7
+OLoss (AONet)	**66.5**	**79.4**	**83.8**	**53.1**

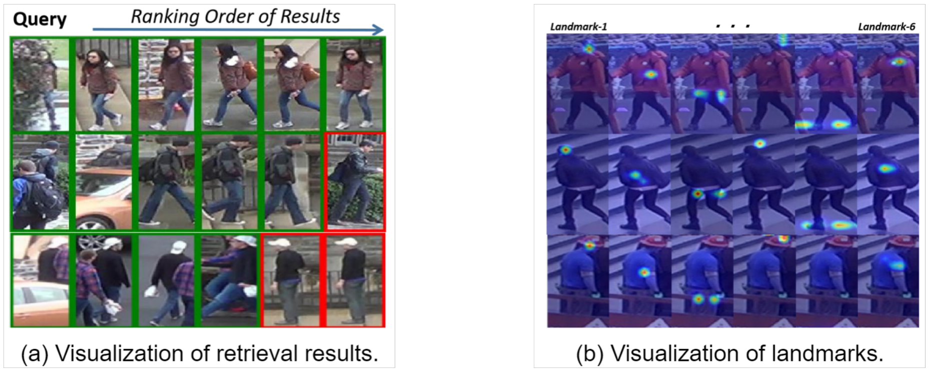
marks. When we replace *SLFea* with *OAFea* that incorporates occlusion awareness, we obtain an improvement on Rank1 (4%) and mAP (6%) (Base_SLFea *vs* AONet†). Meanwhile, the involving of *OAScore* achieves an extra improvement on Rank1 (3%) and mAP (1.2%) (AONet† *vs* AONet‡), which means both *OAScore* and *OAFea* effectively mitigate feature mislocalization caused by occlusion. Furthermore, utilizing *OLoss* does work very effectively with an improvement of 0.9% and 0.6% on Rank1 and mAP respectively.

Comparisons to various baselines. Firstly, we construct three comparable baselines: i) *Base_ Pose* refers to method using external pose features [20] as the pedestrian representation. ii) *Base_ Max* means directly choosing the maximum value in normalized responses without referring to memory. iii) *Base_ RegL* is the method of position regularization loss [27].

As shown in Table 5, the method utilizing landmark features of *OAFea* obviously gains better performance. Thus, it is indeed crucial to involve occlusion awareness in landmark representation. Meanwhile, the performance improves obviously by comprehensively weakening the occluded landmark features, no matter the method with *OAScore* or Base_Max. That is, a reasonable awareness of occlusion indeed brings better performance. However, simply taking the maximum value cannot accurately sense occlusion, but the *OAScore* referring

Table 6. The performance of AONet with different numbers of landmarks on Occluded Person ReID.

Methods	Rank1	Rank5	Rank10	mAP
AONet (K=2)	64.3	79.7	84.5	52.3
AONet (K=4)	65.2	79.0	84.3	53.4
AONet (K=6)	66.5	79.4	83.8	53.1
AONet (K=8)	66.2	79.6	83.9	53.0
AONet (K=10)	65.7	79.8	84.3	52.7



(a) Visualization of retrieval results.

(b) Visualization of landmarks.

Fig. 4. (a) Visualization of retrieval results. The 1st image in each row is the query, and the next five images are returned images with descending ranking. Green and red rectangles indicate correct and error results. (b) Visualization of landmarks in the original pedestrian image. Each column of images refers to the visualization of a specific landmark by its corresponding response map.

to memorized features can effectively handle this problem with 2.2% and 0.8% improvements on both Rank1 and mAP.

Besides, we evaluate the comparable efficiency of *OLoss* by the orthogonal loss and *RegL* with the position regularization loss [27]. As shown in Table 5, the *OLoss* does show obvious improvements on performance. The reason may be that, the position regularization loss, while enabling different landmarks to indicate different regions, does not guarantee attention to select discriminative landmark features, which is however what our orthogonal constraint is good at.

Influence of Number of Landmarks. As shown in Table 6, the performance improves at first as the number of landmarks increases, possibly because more local features provide more robustness to occlusion. However, too many landmarks lead the network to focus on more fine-grained local features, or even background noise, which lacks sufficient discrimination for identification.

Visualization Analysis. We visualize the image retrieval results of the AONet approach in Fig. 4 (a). We get the correct image by AONet for both

Table 7. Comparison of costs with the state-of-the-arts.

Methods	FLOPs(G)	#Params(M)
PGFA[ICCV19]	29.51	57.51
HOReID[CVPR20]	35.80	109.23
SORN[TCSVT20]	24.73	41.96
SGSFA[ACML20]	16.13	47.71
MoS[AAAI21]	12.57	24.22
AONet	6.22	30.22

horizontal and vertical occlusion, as well as to object and pedestrian occlusion. However, when the effective region is too small, the retrieval easily makes mistakes. We also visualize the landmark response map. As shown in Fig. 4 (b), each landmark focuses on a different unique semantic pattern.

Cost Evaluation. To more clearly quantify the advantages of our model over other state-of-the-art models. As shown in Table 7, we compare the number of model parameters "Param" and floating-point operations "FLOPs", where FLOPs are calculated at an input size of 256 × 128. Since the AONet does not use any additional models, such as models of pose estimation and semantic segmentation, it has a smaller time and space overhead. Besides, our AONet has good parallel computing properties while not relying on any additional model, so it is computed at the fastest speed.

5 Conclusion

Previous works for occluded person ReID often rely on external tasks, *e.g.*, pose estimation or semantic segmentation, to extract local features over fixed given regions. In this paper, we propose an end-to-end Attentional Occlusion-aware Network (AONet), including a Landmark Activation layer to extract the landmark features, and an Occlusion Awareness (OA) score to explicitly measure the occlusion. Without any external information by extra tasks, we adaptively extract discriminate anti-occlusion local features with the landmark patterns. The OA is the focus of this paper, on the one hand, providing occlusion reference information to prevent landmark patterns from focusing on the wrong region, and on the other hand, to generate occlusion awareness scores to reduce the weight of the occluded landmark features in classification loss and image matching.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant No. 61972036.

References

1. Chen, D., Xu, D., Li, H., Sebe, N., Wang, X.: Group consistent similarity learning via deep crf for person re-identification. pp. 8649–8658 (2018)
2. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. pp. 8351–8361 (2019)
3. Gao, L., Zhang, H., Gao, Z., Guan, W., Cheng, Z., Wang, M.: Texture semantically aligned with visibility-aware for partial person re-identification. pp. 3771–3779 (2020)
4. Gao, S., Wang, J., Lu, H., Liu, Z.: Pose-guided visible part matching for occluded person reid. pp. 11744–11752 (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. pp. 770–778 (2016)
6. He, L., Liang, J., Li, H., Sun, Z.: Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. pp. 7073–7082 (2018)
7. He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z., Feng, J.: Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. pp. 8450–8459 (2019)
8. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv:1703.07737 (2017)
9. Huang, H., Li, D., Zhang, Z., Chen, X., Huang, K.: Adversarially occluded samples for person re-identification. pp. 5098–5107 (2018)
10. Jia, M., Cheng, X., Zhai, Y., Lu, S., Ma, S., Zhang, J.: Matching on sets: Conquer occluded person re-identification without alignment (2021)
11. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic gradient descent. pp. 1–15 (2015)
12. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person reidentification. pp. 2285–2294 (2018)
13. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. pp. 542–551 (2019)
14. Ren, X., Zhang, D., Bao, X.: Semantic-guided shared feature alignment for occluded person re-identification. pp. 17–32 (2020)
15. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. pp. 17–35 (2016)
16. Sarfraz, M.S., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. pp. 420–429 (2018)
17. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. pp. 3960–3969 (2017)
18. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. pp. 1229–1240 (2018)
19. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Part-aligned bilinear representations for person re-identification. pp. 402–419 (2018)
20. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. pp. 5693–5703 (2019)
21. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. pp. 393–402 (2019)
22. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. pp. 393–402 (2019)

23. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). pp. 480–496 (2018)
24. Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J., Yan, J., Wang, X.: Eliminating background-bias for robust person re-identification. pp. 5794–5803 (2018)
25. Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J.: High-order information matters: Learning relation and topology for occluded person re-identification. pp. 6449–6458 (2020)
26. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. pp. 274–282 (2018)
27. Xie, W., Shen, L., Zisserman, A.: Comparator networks. pp. 782–797 (2018)
28. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. pp. 2119–2128 (2018)
29. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. pp. 1389–1398 (2019)
30. Yao, H., Zhang, S., Hong, R., Zhang, Y., Xu, C., Tian, Q.: Deep representation learning with part loss for person re-identification **28**(6), 2860–2871 (2019)
31. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
32. Yu, Q., Chang, X., Song, Y.Z., Xiang, T., Hospedales, T.M.: The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching (2018)
33. Zhang, X., Yan, Y., Xue, J.H., Hua, Y., Wang, H.: Semantic-aware occlusion-robust network for occluded person re-identification **31**(7), 2764–2778 (2020)
34. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. pp. 3219–3228 (2017)
35. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. pp. 1116–1124 (2015)
36. Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S.: Partial person re-identification. pp. 4678–4686 (2015)
37. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. pp. 1318–1327 (2017)
38. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. vol. 34, pp. 13001–13008 (2020)