

# Quality-aware Part Models for Occluded Person Re-identification

Pengfei Wang, *Student Member, IEEE*, Changxing Ding, *Member, IEEE*, Zhiyin Shao, Zhibin Hong, Shengli Zhang, *Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

**Abstract**—Occlusion poses a major challenge for person re-identification (ReID). Existing approaches typically rely on outside tools to infer visible body parts, which may be suboptimal in terms of both computational efficiency and ReID accuracy. In particular, they may fail when facing complex occlusions, such as those between pedestrians. Accordingly, in this paper, we propose a novel method named **Quality-aware Part Models (QPM)** for occlusion-robust ReID. First, we propose to jointly learn part features and predict part quality scores. As no quality annotation is available, we introduce a strategy that automatically assigns low scores to occluded body parts, thereby weakening the impact of occluded body parts on ReID results. Second, based on the predicted part quality scores, we propose a novel **identity-aware spatial attention (ISA) module**. In this module, a coarse identity-aware feature is utilized to highlight pixels of the target pedestrian, so as to handle the occlusion between pedestrians. Third, we design an adaptive and efficient approach for generating global features from common non-occluded regions with respect to each image pair. This design is crucial, but is often ignored by existing methods. QPM has three key advantages: 1) it does not rely on any outside tools in either the training or inference stages; 2) it handles occlusions caused by both objects and other pedestrians; 3) it is highly computationally efficient. Experimental results on four popular databases for occluded ReID demonstrate that QPM consistently outperforms state-of-the-art methods by significant margins. The code of QPM will be released.

**Index Terms**—Person Re-identification, Occlusion, Attention Models

## I. INTRODUCTION

**P**ERSON re-identification (ReID) involves spotting a specific person of interest, e.g. a missing child, across disjoint camera views. Due to the widespread deployment of modern surveillance systems, ReID has attracted increasing attention from both academia and industry [1]–[13]. Most existing ReID approaches assume that the pedestrian’s entire body is visible, and tend to ignore the more challenging occlusion situations. However, in real-world applications, pedestrians are very often occluded by objects or other pedestrians.

Occlusion poses a major challenge for ReID, as it affects the appearance of pedestrian. As illustrated in Fig. 1(a), similar

Pengfei Wang, Changxing Ding, and Zhiyin Shao are with the School of Electronic and Information Engineering, South China University of Technology, 381 Wushan Road, Tianhe District, Guangzhou 510000, P.R. China (e-mail: eepengfei.wang@mail.scut.edu.cn; chxding@scut.edu.cn; eezysao@mail.scut.edu.cn).

Zhibin Hong is with department of Computer Vision Technology (VIS), Baidu Inc, ShenZhen 518000, China (e-mail: hongzhibin@baidu.com).

Shengli Zhang is with the College of Electronic and Information Engineering, Shenzhen University, Shenzhen 518052, China (e-mail: zsl@szu.edu.cn).

Dacheng Tao is with JD Explore Academy, JD.com, Beijing 100176, China (e-mail: taodacheng@jd.com).



Fig. 1. Example images that illustrate the challenges for occluded ReID: (a) Similar occlusion reduces inter-class distance. (b) Dissimilar occlusion enlarges intra-class variation. (c) Both human parsing and pose estimation tools may fail when facing complex occlusions, e.g. occlusion between pedestrians. (d) Both tools may ignore discriminative accessories, such as backpacks or handbags. (Best viewed in color.)

occlusion reduces inter-class distance, which indicates that images of different identities may have similar visual features. Moreover, as shown in Fig. 1(b), different occlusions enlarge intra-class distance, meaning that two images of the same pedestrian may be quite different in terms of their appearance. This is because occlusions may differ as regards their location and content; therefore, occlusion tends to result in incorrect retrieval results.

Since occlusion changes pedestrian appearance, one intuitive solution is to only utilize visible body parts for ReID. Most existing methods adopt this strategy [14]–[16]. However, they usually rely on **outside tools** to acquire the visibility cues of body parts (e.g. the prediction confidence of pose estimation models). Beside extra computational cost, this strategy may not be robust to complex occlusions, such as occlusions between pedestrians. As illustrated in Fig. 1(c), both human parsing and pose estimation tools may fail when facing occlusions between pedestrians. Moreover, visibility is not equivalent to discriminative power. On the one hand, one visible part may look quite similar across different pedestrians. For example, nearly all pedestrians’ forearms are not covered by any clothes in images captured in summer. On the other hand, invisible body parts may be occluded by discriminative accessories, such as backpacks and bags, as shown in Fig. 1(d). These accessories are critical for ReID, but tend to be ignored by outside tools. It is therefore reasonable to seek out a robust and easy-to-use method that automatically infers and utilizes discriminative body parts to handle the occlusion problem.

Accordingly, in this paper, we propose a novel framework named **QPM** for occluded ReID. QPM includes a part branch and a global branch. The part branch automatically infers **part-specific quality scores** rather than visibility scores. More specifically, it jointly learns discriminative part features and predicts part quality scores in an end-to-end fashion. This is achieved by including both pair-wise part distances and pair-wise part quality scores in the triplet loss [17]; as a result, the part branch can automatically assign low quality scores to poor-quality body parts in order to weaken their influence. Another key benefit of our approach is that it is independent from any outside tools and does not require annotations of quality scores. However, when a pedestrian is occluded by other persons, the model may predict a relatively high score for occluded regions; this is because the model cannot well differentiate body parts of different pedestrians. We solve this problem in the global branch.

The global branch includes two main components. First, we propose an identity-aware spatial attention (ISA) module based on the predicted part quality scores. In this module, a coarse identity-aware feature is processed by a simple two-layer network and optimized using the cross-entropy loss function. Then, it is utilized to suppress noisy responses and highlight responses from the body region of the pedestrian needing to be identified. Therefore, it can be used to handle occlusions between pedestrians. Second, we design an adaptive and efficient approach to generate global features from the common non-occluded regions with respect to each image pair. By contrast, existing works typically extract fixed global features, ignoring the difference in occlusion locations between a pair of images.

In the inference stage, both the part and global features are utilized to calculate the similarity score between each pair of images. The weighted average of both scores represents the overall similarity of an image pair. We conduct extensive experiments on four popular datasets for occluded ReID, i.e. Partial-iLIDS [18], Partial-REID [19], Occluded-Duke [14], and P-DukeMTMC [20]. The results show that our simple QPM model consistently outperforms existing approaches by significant margins. Moreover, our approach enjoys further advantages of being robust and easy to use.

In conclusion, the main contributions of this paper are summarized as following:

- We propose an end-to-end framework that jointly learns discriminative features and predicts part quality scores. Compared with existing works, it does not rely on any outside tools in either the training or inference stages.
- We propose a novel identity-aware spatial attention (ISA) approach that efficiently handles the occlusion between pedestrians. Experimental results prove that it outperforms existing spatial attention methods.
- We introduce an Adaptive Global Feature Extraction (AGFE) module that extracts global features from the commonly non-occluded regions for each image pair, which significantly promotes ReID performance.

The remainder of this paper is organized as follows. We first review the related works in Section II. Then, we describe the proposed QPM in more detail in Section III. Extensive experimental results on three benchmarks are reported and

analyzed in Section IV, after which the conclusions of the present work are outlined in Section V.

## II. RELATED WORK

### A. Occluded Person ReID Models

One main challenge for occluded ReID is to identify visible body regions. Most existing works utilize visibility cues provided by outside tools [6], [21]–[25]. Here, we divide occluded ReID methods into two categories depending on whether or not outside tools are required during training and testing.

The first category of methods employs outside tools in both the training and testing stages [14]–[16]. For example, Miao *et al.* [14] utilized pose landmarks to identify visible local patches and only adopt commonly visible local patches of one image pair for matching. Wang *et al.* [15] utilized pose landmarks to learn high-order relation and topology information of the visible local features, so as to better match probe with gallery images. Gao *et al.* [16] employed graph matching and utilized pose landmarks to self-mine part visibility scores. They then match probe and gallery images by calculating the part-to-part distances in visible regions. However, in addition to its extra computational cost, another key downside of this approach is that external tools may not be reliable when encountering complex occlusions, as illustrated in Fig. 1(c-d).

The second category of methods avoids using outside tools in the testing stage [26]–[29]. For example, under the guidance of human masks, He *et al.* [26] designed an occlusion-sensitive foreground probability generator that enables the model to focus on non-occluded human body parts. He *et al.* [27] further combined pose landmarks and human masks to generate spatial attention maps that guide discriminative feature learning. These approaches can reduce the impact of occlusions on the extracted features. Despite the convenience this affords during testing, this approach still relies on the visibility information of body part for each training image.

Moreover, two very recent works [30], [31] try to predict visible scores without outside tools. VPM [31] classifies each pixel in the image into one body part. However, it is designed for the partial ReID problem and cannot be directly used to solve the occluded ReID problem. Specifically, VPM classifies each pixel in the partial image into one body part, assuming there is no occlusion in the partial image. ISP [30] performs cascaded clustering on feature maps to generate the pseudo-labels of human parts for each pixel. However, ISP contains no strategy to handle the occlusion problem between pedestrians. In addition, the clustering process is time-consuming.

In a departure from existing works, our approach aims to predict the part quality rather than visibility in an end-to-end framework. And we do not employ any outside tools or require any quality annotations in either the training or inference stages. Moreover, our approach is robust to complex occlusions. In conclusion, QPM is a powerful and efficient model.

### B. Part-based Person ReID Models

Due to their powerful representation ability, part-based methods are popular for ReID. Depending on the way to

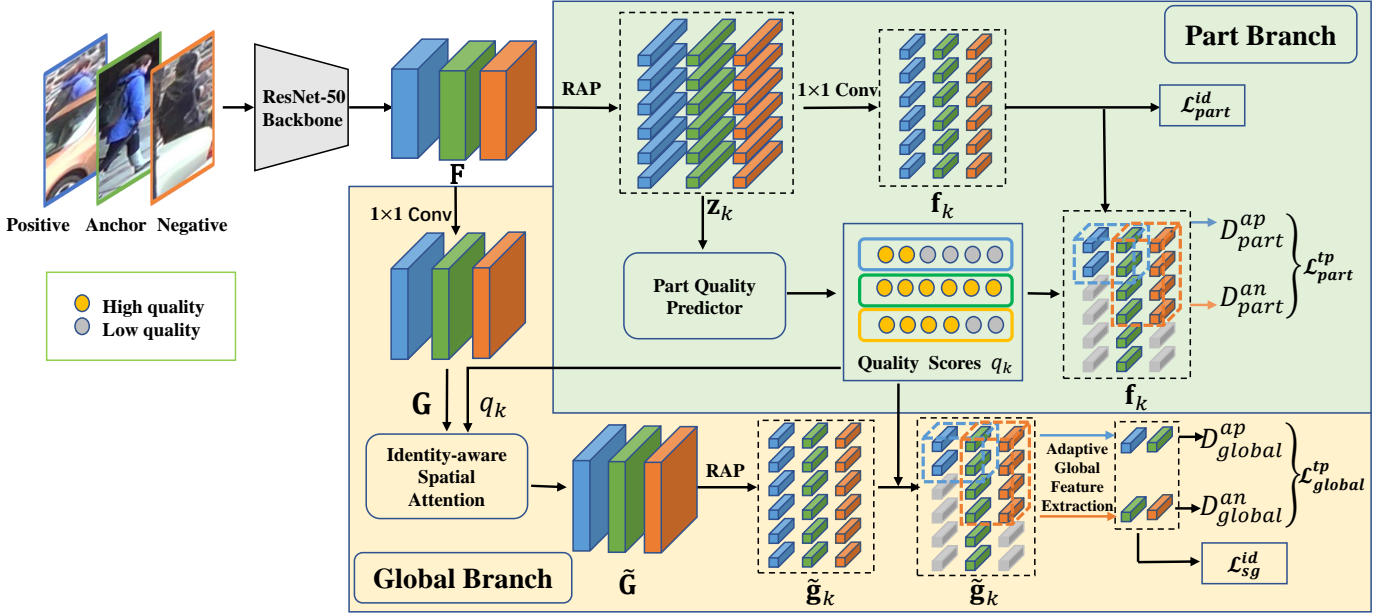


Fig. 2. Model architecture of QPM in the training stage. Based on ResNet-50, QPM builds a part branch and a global branch. The part branch jointly extracts part features and predicts part quality scores. The quality scores are utilized to weaken the impact of occluded body parts. The global branch first adopts a novel Identity-aware Spatial Attention module to handle occlusions between pedestrians. Then, it adaptively and efficiently extracts global features from the common non-occluded regions for each image pair. In the testing stage, the global features, part features, and part quality scores are utilized together for occluded ReID. (Best viewed in color.)

obtain body part locations, we divide existing works into three categories.

**Fixed Location-based Methods.** Methods in this category typically split the output feature maps of one backbone model into several stripes in fixed locations [3], [32], [33]. Part features are then respectively extracted from the stripes. For example, Sun *et al.* [3] uniformly divide the output feature maps into 6 horizontal stripes to represent different part-level features. Wang *et al.* [32] also partition one image into horizontal stripes. The main advantage of this strategy lies in its efficiency.

**Outside Tools-based Methods.** These methods utilize outside tools, e.g. pose estimation [23], [25], [34], [35] and human parsing models [6], [21], to detect body parts. They then extract part features from the detected body parts. There are two key downsides of these approaches: first, they require additional computational cost; second, ReID performance is vulnerable to the reliability of outside tools.

**Attention-based Methods.** These methods predict body part locations based on the feature maps produced by ReID models [2], [36]–[38]. For example, Zhao *et al.* [36] proposed to predict a set of masks and perform element-wise multiplication between one mask and each channel of the feature maps to produce part-specific features. In comparison, Li *et al.* [2] designed a hard regional attention model that can predict bounding boxes for each body part. However, the lack of explicit supervision for part alignment may cause difficulty in the optimization of attention models.

In part-based ReID methods, it is a common practice to concatenate part features as the final representation [3], [23], [33], [36], [39]–[42]. However, it is less effective for occluded ReID, as it ignores the impact of features from occluded parts.

In this paper, we accordingly propose to jointly learn part features and predict part quality. For simplicity, we extract part features from fixed part locations. If part locations are provided by outside tools or attention modules, the performance of our approach can be further promoted.

### III. METHODS

The architecture of QPM is illustrated in Fig. 2. It consists of a **part feature learning branch** and a **global feature learning branch**. The part branch outputs  $K$  part-level features, as well as  $K$  quality scores that indicate the discriminative power of each body part. The global branch generates global features that are adaptively and efficiently extracted from the common non-occluded regions for each gallery-query image pair. In the following, we will introduce three key designs in the two branches individually.

#### A. Joint Learning Part Feature and Quality Scores

**Part Feature Extractor.** Following [3], we adopt ResNet-50 as backbone and remove its last spatial down-sampling operation to increase the size of the output feature maps. The output feature maps are denoted as  $F$  for simplicity. To obtain the part features,  $F$  is first uniformly split into  $K$  parts in the vertical orientation. Following [3], we set  $K$  as 6 in this work. Next, the feature maps for each part are processed by a Region Average Pooling (RAP) operation and one  $1 \times 1$  Conv layer. The parameters of the Conv layer are not shared between parts. For the  $k$ -th part, the feature vectors before and after the  $1 \times 1$  Conv layer are denoted as  $z_k$  and  $f_k$ , respectively.



$\mathbf{f}_k$  is utilized as the final part feature and is optimized by the cross-entropy loss:

$$\mathcal{L}_{part}^{id} = \frac{1}{N} \sum_{l=1}^N \sum_{k=1}^K \mathcal{L}_{ce}(\mathbf{W}_k^p \mathbf{f}_k^l), \quad (1)$$

where  $N$  denotes the batch size.  $\mathbf{f}_k^l \in \mathbb{R}^d$  represents the  $k$ -th part-level feature for the  $l$ -th image in a batch.  $\mathbf{W}_k^p$  stands for the parameters of the classification layer for the  $k$ -th part feature.  $\mathcal{L}_{ce}$  stands for the cross-entropy loss function.

**Part Quality Predictor.** A key challenge for occluded ReID is to identify visible body parts. Recent works [14], [16] typically rely on outside tools to infer whether or not one part is visible. However, as argued in Section I, visibility is not equivalent to discriminative power. We accordingly propose an efficient method to predict part feature quality rather than part visibility. As shown in Fig. 2, we feed  $\mathbf{z}_k$  into the part quality predictor module. This module comprises one  $1 \times 1$  Conv layer, one batch normalization (BN) [43] layer, and one sigmoid activation layer. The output of the sigmoid layer is the quality score of the  $k$ -th part. The parameters of the predictor are not shared between parts.

#### Joint Optimization.

As there is no annotation of part quality scores provided, we cannot impose a direct supervision signal on the predicted part quality. Recent works for video-based recognition have revealed that framewise features and quality scores can be jointly optimized with the same identification [44], [45] or metric learning loss [46]. In these works, e.g., QAN [46] and MG-RAFA [44], the quality scores are utilized to aggregate multiple frame-level features into a single video-level feature. However, this approach cannot be directly used in image-based ReID, as recognition is based on each single image. Inspired by these works, we propose to jointly optimize part features and part quality scores using triplet loss. Accordingly, the quality scores in QPM are imposed on part distances instead of frame features. Moreover, both pair-wise part distances and pair part quality scores are included in the triplet loss, instead of using a sample's own quality score only. In this way, our approach optimizes pair-wise part distance and pair part quality scores for occluded ReID in an end-to-end manner.

Specifically, given a probe image  $I^p$  and a gallery image  $I^g$  to be compared, we first calculate their part-wise cosine distances  $d_k^{pg}$  ( $1 \leq k \leq K$ ). Next, the part-wise distances are summed via weighted average, as follows:

$$D_{part}^{pg} = \frac{\sum_{k=1}^K q_k^p q_k^g d_k^{pg}}{\sum_{k=1}^K q_k^p q_k^g}, \quad (2)$$

where  $q_k^p$  and  $q_k^g$  are the quality score of  $k$ -th part for probe image  $I^p$  and gallery image  $I^g$ , respectively. In this way, body parts with low quality scores will contribute less to  $D_{part}^{pg}$ , which weakens the impact of occlusion.

To sample sufficient triplets during training, we randomly sample  $A$  images in each of  $P$  random identities to create a

mini-batch, the batchsize  $N$  of which is equal to  $P \times A$ . The triplet loss is formulated as  $\mathcal{L}_{part}^{tp} =$

$$\frac{1}{N_{tp}} \sum_{i=1}^P \sum_{a=1}^A [\alpha + \max_{p=1 \dots A} D_{part}^{a_i p_i} - \min_{\substack{n=1 \dots A \\ j \neq i}} D_{part}^{a_i n_j}]_+, \quad (3)$$

where  $\alpha$  is the margin of the triplet constraint, while  $N_{tp}$  is the number of triplets in a batch that violate the triplet constraint and  $[\cdot]_+$  denotes the hinge loss.  $D_{part}^{a_i p_i}$  and  $D_{part}^{a_i n_j}$  are calculated by Eq. 2 and denote the distances of the positive and negative image pairs in a triplet, respectively. In order to reduce triplet loss, the part quality predictors have to predict lower scores to occluded parts.

**Discussion.** With the help of the above constraints, the model can predict the quality scores of the part features without using any external tools. However, this approach still has limitations. Take the predicted quality scores in Fig. 4 as an example. It can be seen that the model works well for occlusions caused by objects (e.g. cars, trees, and boxes). In these situations, the quality scores of occluded body parts are very low. However, when a pedestrian is occluded by other persons, the model may predict a relatively high score for occluded regions; this is because the model cannot well differentiate body parts of different pedestrians. In the following, we handle the above problem in the global feature learning branch. Most existing approaches [14], [15], [29] simply extract global features from visible regions for each image. However, this strategy suffers from two problems. First, as explained above, visibility or quality prediction of image regions may be interfered by occlusions between pedestrians. Second, as shown in Fig. 1(b), two images may differ in occlusion locations, meaning that they are not directly comparable even if features are extracted from visible regions for each image. In the following, we propose an Identity-aware Spatial Attention (ISA) approach and an adaptive global feature extraction approach to handle each of these two problems, respectively.

#### B. Identity-aware Spatial Attention

Fig. 3 illustrates the structure of ISA. ISA makes use of a coarse identity-aware feature to generate spatial attention for each image. This attention suppresses occlusions caused by objects and other pedestrians, meaning that only features of spatial regions relevant to the target pedestrian are highlighted.

In more detail, we feed  $\mathbf{F}$  into another  $1 \times 1$  Conv layer and obtain the feature maps  $\mathbf{G}$ . Global features are then extracted based on  $\mathbf{G}$ . Similar to the part branch, we uniformly partition  $\mathbf{G}$  into  $K$  parts and obtain  $K$  part-level features that are denoted as  $\mathbf{g}_k$  ( $1 \leq k \leq K$ ) via RAP operations. We then obtain a coarse identity-aware global feature  $\mathbf{h}$  by fusing  $\mathbf{g}_k$  via weighted averaging as follows:

$$\mathbf{h} = \sum_{k=1}^K \hat{q}_k \mathbf{g}_k, \quad (4)$$

where  $\hat{q}_k$  is the normalized part quality score produced in the part branch in Section III-A. Formally,

$$\hat{q}_k = \frac{q_k}{\sum_{i=1}^K q_i}. \quad (5)$$

To further reduce the impact of occlusions, we process  $\mathbf{h}$  using a simple two-layer network inspired by the squeeze-and-excitation network [47]. As illustrated in Fig. 3, the dimension of  $\mathbf{h}$  is first reduced via a  $1 \times 1$  Conv layer which is followed by a ReLU layer, then recovered to the original dimension by means of another  $1 \times 1$  Conv layer. The reduction ratio of the first Conv layer is set to 4; the output of the two Conv layers are denoted as  $\hat{\mathbf{h}}$  and  $\tilde{\mathbf{h}}$ , respectively.

Moreover, to ensure that noisy elements in  $\mathbf{h}$  are suppressed via the reduction operation, we impose a cross-entropy loss on  $\hat{\mathbf{h}}$  as follows:

$$\mathcal{L}_{global}^{id} = \frac{1}{N} \sum_{l=1}^N \mathcal{L}_{ce}(\mathbf{W}^g \hat{\mathbf{h}}^l), \quad (6)$$

where  $\hat{\mathbf{h}}^l$  represents the feature produced by the first Conv layer for the  $l$ -th image in a batch, while  $\mathbf{W}^g$  denotes the parameters of the classification layers.

By adopting this approach, information in  $\tilde{\mathbf{h}}$  is identity-aware and noise-free. Accordingly, we employ  $\tilde{\mathbf{h}}$  to generate the spatial attention map  $\mathbf{M}$  for the feature maps  $\mathbf{G}$ . Formally:

$$\mathbf{M} = \sigma(\mathbf{G} * \tilde{\mathbf{h}}), \quad (7)$$

where  $\sigma$  is a sigmoid function, while  $*$  represents the inner product between  $\tilde{\mathbf{h}}$  and the feature vector of each pixel in  $\mathbf{G}$ . Accordingly,  $\mathbf{M}$  is a matrix with the same height and width as  $\mathbf{G}$ . As identity-relevant pixels obtain a high response value in  $\mathbf{M}$ , we apply  $\tilde{\mathbf{M}}$  to weigh  $\mathbf{G}$  and produce new feature maps denoted as  $\tilde{\mathbf{G}}$ . The above process can be summarized as follows:

$$\tilde{\mathbf{G}} = \mathbf{M} \odot \mathbf{G} + \mathbf{G}, \quad (8)$$

where  $\odot$  signifies the element-wise multiplication between  $\mathbf{M}$  and each channel of  $\mathbf{G}$ .

**Discussion.** To the best of our knowledge, ISA is one of the first efficient method to address occlusions between pedestrians in occluded ReID. A few most recent works [48], [49] can potentially solve this problem. These approaches adopt co-attention mechanism and attempt to search for pixel-level correspondence between each pair of images, enabling features to be extracted from semantically corresponding regions. However, during the training and inference stage, these approaches adopt computationally expensive matrix multiplication to infer semantically corresponding pixels for each pair of query and gallery images. Obviously their computationally cost is significantly higher than that of ISA. Therefore, our method has obvious advantages in efficiency for ReID.

### C. Adaptive Global Feature Extraction

As indicated in Fig. 5, the responses on  $\tilde{\mathbf{G}}$  focus primarily on the body of the pedestrian to identify after the processing of ISA. However, this does not mean that it is reasonable to extract global-level features directly from  $\tilde{\mathbf{G}}$ ; this is because the two images being compared may differ in terms of their occlusion locations. To ensure semantic consistency, it is essential to adaptively extract global features from the common non-occluded regions for each image pair. We designed the

Adaptive Global Feature Extraction (AGFE) module with the help of the part quality score to achieve this goal.

For example, given a probe image  $I^p$  and a gallery image  $I^g$ , we first obtain their feature maps  $\tilde{\mathbf{G}}^p$  and  $\tilde{\mathbf{G}}^g$  from the output of the ISA module. In the next step, we equally partition each of them into  $K$  parts and apply the RAP operation on each divided feature maps. In this way, we obtain a set of  $K$  feature vectors for  $I^p$  and  $I^g$ , denoted as  $\tilde{\mathbf{g}}_k^p$  and  $\tilde{\mathbf{g}}_k^g$  respectively. We then adopt the part quality scores from the part branch to aggregate  $\tilde{\mathbf{g}}_k^p$  and  $\tilde{\mathbf{g}}_k^g$  and obtain the global-level features  $\mathbf{h}_g^p$  and  $\mathbf{h}_g^g$  for  $I^p$  and  $I^g$ , respectively. More specifically,

$$\mathbf{h}_g^p = \sum_{k=1}^K \tilde{q}_k \tilde{\mathbf{g}}_k^p, \quad (9)$$

$$\mathbf{h}_g^g = \sum_{k=1}^K \tilde{q}_k \tilde{\mathbf{g}}_k^g, \quad (10)$$

where  $\tilde{q}_k$  denotes the weight for  $\tilde{\mathbf{g}}_k^p$  and  $\tilde{\mathbf{g}}_k^g$ .  $\tilde{q}_k$  is computed as follows:

$$\tilde{q}_k = \frac{q_k^p q_k^g}{\sum_{i=1}^K q_i^p q_i^g}. \quad (11)$$

The classification loss for the final global representations can thus be formulated as follows:

$$\mathcal{L}_{sg}^{id} = \frac{1}{NN} \sum_{g=1}^N \sum_{p=1}^N (\mathcal{L}_{ce}(\mathbf{W}^s \mathbf{h}_g^p) + \mathcal{L}_{ce}(\mathbf{W}^s \mathbf{h}_g^g)). \quad (12)$$

We also apply the triplet loss to ensure that the intra-class distances are smaller than the inter-class distances. This triplet loss is similar to Eq. 3:

$$\mathcal{L}_{global}^{tp} = \frac{1}{N_{tp}} \sum_{i=1}^P \sum_{a=1}^A [\alpha + \max_{p=1 \dots A} D_{global}^{a;p_i} - \min_{\substack{n=1 \dots A \\ j=1 \dots P \\ j \neq i}} D_{global}^{a;n_j}]_+, \quad (13)$$

where  $D_{global}^{a;p_i}$  and  $D_{global}^{a;n_j}$  represent the cosine distances of global features which are obtained by Eq. 9 and Eq. 10 for the positive and negative image pairs in a triplet, respectively.

Obtaining global features from the common non-occluded regions of each image pair is usually ignored by existing works. As shown in Table V, the AGFE module significantly promotes performance for occluded ReID. This is because the AGFE module extracts semantically aligned global features.

### D. Occluded ReID via QPM

During training, the overall objective function of QPM can be written as follows:

$$\mathcal{L} = \mathcal{L}_{part}^{id} + \mathcal{L}_{part}^{tp} + \mathcal{L}_{global}^{id} + \mathcal{L}_{sg}^{id} + \mathcal{L}_{global}^{tp}. \quad (14)$$

During the training process, the overall loss is optimized together. The parameters of the network, including those of the fully connected layers, are optimized together via gradient descent.

Many existing works design local and global branches [14], [15], [50], [51] for multi-source structural information integration. Similar to [14], there are two parts in QPM that make up the final distance between one pair of query

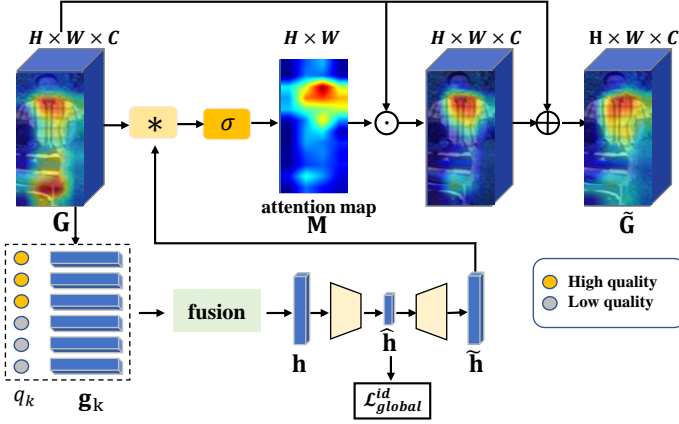


Fig. 3. Structure of the Identity-aware Spatial Attention (ISA) module. It first adopts a coarse identity-aware global feature vector  $\mathbf{h}$  and a lightweight two-layer network to generate a spatial attention map  $\mathbf{M}$ .  $\mathbf{M}$  is applied on feature maps, i.e.  $\mathbf{G}$ , to suppress responses of occluded pixels. As  $\mathbf{M}$  is identity aware, occlusions between pedestrians can be suppressed. (Best viewed in color.)

and gallery images: namely, the distance between part-level features and the distance between global-level features. In our approach, the distance between the part-level features is computed according to Eq. 2. The global-level features with respect to the image pair are obtained according to Eq. 9 and Eq. 10. Formally,

$$D^{pq} = \gamma D_{part}^{pq} + (1 - \gamma) D_{global}^{pq}, \quad (15)$$

where  $\gamma$  is the weight that balances the contributions from  $D_{part}^{pq}$  and  $D_{global}^{pq}$ .  $\gamma$  is consistently set to 0.6 in this work.

In the inference stage, we first compute the distance between a query image and each of the gallery images using the part features according to Eq. 2. The body parts with low quality scores will contribute less to the distance, which weakens the impact of occlusion. In this way, we efficiently obtain the top  $n$  nearest neighbors for the query image. Then, we compute the final distance according to Eq. 15 between the query image and each of the  $n$  nearest neighbors. Therefore, the AGFE module hardly increases the inference time cost.

## IV. EXPERIMENTS

### A. Datasets and Settings

**Datasets.** We conduct experiments on four popular databases for occluded ReID, i.e. Partial-iLIDS [18], Partial-REID [19], Occluded-Duke [14] and P-DukeMTMC [20].

**Partial-iLIDS** [18] was constructed based on the iLIDS [52] dataset. It contains 238 images of 119 identities, all of which were captured in an airport. Some images in the dataset contain people occluded by other individuals or luggages. Each pedestrian has 1 full-body image and 1 occluded image. All probe images are occluded person images, while all gallery images are holistic images.

**Partial-REID** [19] was collected at a university campus and includes 600 images of 60 pedestrians. Each person has 5 full-body images and 5 occluded images. These images are collected from different viewpoints, backgrounds, and different

TABLE I  
PERFORMANCE COMPARISONS ON OCCLUDED-DUKE UNDER SUPERVISED SETTING. \* INDICATES THAT A DIFFERENT BACKBONE IS USED. + REPRESENTS THE EXTENDED VERSION OF THE CONFERENCE PAPER.

Method	Rank-1	Rank-5	Rank-10	mAP
Part-Aligned [36]	28.8	44.6	51.0	20.2
Part-Bilinear [56]	36.9	-	-	-
Random Erasing [57]	40.5	59.6	66.8	30.0
DSR [18]	40.8	58.2	65.2	30.4
SFR [28]	42.3	60.3	67.3	32.0
PCB [3]	42.6	57.1	62.9	33.7
Adver Occluded [58]	44.5	-	-	32.2
PVPM [16]	47.0	-	-	37.7
PGFA [14]	51.4	68.6	74.9	37.3
HOREID [15]	55.1	-	-	43.8
PGFA+ [59]	56.3	72.4	78.0	43.5
ISP* [30]	62.8	78.1	82.9	52.3
<b>QPM</b>	<b>64.4</b>	<b>79.3</b>	<b>84.2</b>	<b>49.7</b>
QPM*	<b>66.7</b>	<b>80.2</b>	<b>84.4</b>	<b>53.3</b>

types of severe occlusion. All probe images are occluded person images, while all gallery images are holistic images.

**Occluded-Duke** [14] was constructed based on the DukeMTMC [53] database. It is composed of 15,618 training images of 702 identities, 2,210 occluded query images of 519 identities, and 17,661 gallery images. There are rich variations in Occluded-Duke, including different viewpoints and a large variety of obstacles, including cars, bicycles, trees, other persons. Occluded-Duke is a more difficult and practical dataset since both probe and gallery images have occlusions.

**P-DukeMTMC** [20] is another subset of DukeMTMC [53]. There are 12,927 training images of 665 identities, 2,163 query images of 634 identities, and 9,053 gallery images. These images are occluded by different types of occlusion in public, e.g., people, luggages, cars and guideboards.

**Implementation Details.** We conduct experiments using the Pytorch framework. We set both  $P$  and  $A$  to 8; therefore, the batch size is 64. We adopt random erasing to simulate occlusion. All images are resized to  $384 \times 128$  pixels and augmented via random horizontal flipping. The number of body parts, i.e.  $K$ , is set to 6. The margin  $\alpha$  for the triplet loss is set to 0.3. The number of the nearest neighbors, i.e.,  $n$ , is set as 30. The SGD optimizer is utilized for model optimization. Following [3], [32], [41], we do not use weight regularization in the SGD optimizer. Fine-tuned from the IDE model [54], the QPM is trained in an end-to-end fashion for 70 epochs. The initialized learning rate is set to 0.01 and is reduced by multiplying 0.1 for every 20 epochs.

**Evaluation Protocols.** We report the Cumulated Matching Characteristics (CMC) and mean Average Precision (mAP) value for the proposed approach. The evaluation package is provided by [55], and all the experimental results are obtained in the single query setting.

Moreover, we provide stability analysis on the performance of QPM in the supplementary material.

### B. Performance under Supervised Setting

For the two large-scale datasets, i.e. Occluded-Duke [14] and P-DukeMTMC [20], we train QPM using their own training sets, respectively.

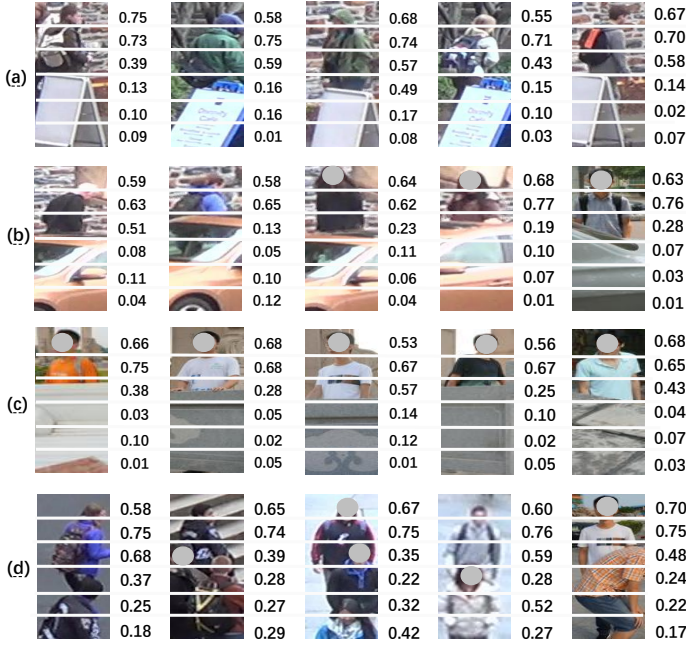


Fig. 4. Example images with predicted part quality scores. (a) Occlusion caused by billboards. (b) Occlusion caused by cars. (c) Occlusion caused by buildings. (d) Occlusion caused by pedestrians.

TABLE II  
PERFORMANCE COMPARISONS ON P-DUKEMTMC UNDER SUPERVISED SETTING.

Method	Rank-1	Rank-5	Rank-10	mAP
Teacher-S [29]	51.4	50.9	-	-
PCB [3]	79.4	87.1	90.0	63.9
IDE [54]	82.9	89.4	91.5	65.9
PVPM [16]	85.1	91.3	93.3	69.9
PGFA <sup>+</sup> [59]	85.7	92.0	94.2	72.4
ISP* [30]	89.0	94.1	95.3	74.7
<b>QPM</b>	<b>89.4</b>	<b>93.9</b>	<b>95.6</b>	<b>74.4</b>
<b>QPM*</b>	<b>90.7</b>	<b>94.4</b>	<b>95.9</b>	<b>75.3</b>

**Results on Occluded-Duke.** The performance of QPM and state-of-the-art methods on Occluded-Duke are tabulated in Table I. Some recent methods have achieved competitive performance with the help of pose landmarks: for example, HOREID [15] learns high-order relation and topology information for local features of visible landmarks, facilitating better match between probe and gallery images, and achieves 55.1% Rank-1 accuracy and 43.8% mAP. In comparison, QPM significantly outperforms HOREID by 9.3% and 5.9% in terms of Rank-1 accuracy and mAP, respectively. Moreover, QPM does not depend on pose estimation tools in either training or testing. This remarkable performance improvement clearly demonstrates the effectiveness of QPM.

One recent method ISP\* [30] achieves strong performance with a deeper backbone model, i.e. HRNet-W32 [60]. With the same backbone model, QPM\* achieves higher Rank-1 and mAP performance than ISP\*.

**Results on P-DukeMTMC.** Comparison results on P-DukeMTMC are summarized in Table II. As the table shows, QPM achieves 89.4% Rank-1 accuracy and 74.4% mAP, surpassing the best previous method PVPM [16] by 4.3% and

4.5% in terms of Rank-1 accuracy and mAP, respectively. The above comparison results are consistent with those obtained on the Occluded-Duke database. These experimental results further demonstrate that our method can effectively solve the occlusion problem for ReID.

### C. Performance under Transfer Setting

Following [15], [16], [18], [31], ReID model in this setting are trained using the Market-1501 database [61]. Then, the model is directly evaluated on the Partial-REID, Partial-iLIDS, and P-DukeMTMC databases.

**Results on Partial-REID and Partial-iLIDS.** Each of the two databases contain two types of testing data: namely, partial images from which occluded regions have been manually removed, and the original occluded images. Similarly, depending on the testing data used, existing methods can be roughly divided into two categories: those using partial images and those using the original occluded images. The performance of QPM and state-of-the-art methods are tabulated in Table III.

As is evident from the table, for Partial-iLIDS, it can be seen that QPM outperforms all other methods that also evaluate on the original occluded images. For example, QPM beats PGFA [14] by 8.2% and 4.8% in terms of Rank-1 and Rank-3 accuracy respectively. It also outperforms one of the most recent methods using partial data, i.e. HOREID [15], by 4.7% in terms of Rank-1 accuracy. For Partial-REID, QPM outperforms all other methods using the original occluded images: for example, QPM beats PVPM [16] by 3.4% in terms of Rank-1 accuracy. While its performance is slightly lower than HOREID [15] (which evaluates on partial data), it should be noted that unlike HOREID, our method does not require either the manual removal of occluded regions during testing or any additional tools. ISP\* [30] achieves 65.7% Rank-1 and 75.3% Rank-3 accuracy on Partial-REID, which is lower than those by QPM. This is because the image quality of Partial-REID is poor and the image resolution is relatively low. In this case, ISP is prone to errors in pixel classification. In comparison, our method is more robust.

**Results on P-DukeMTMC.** Finally, we evaluate the performance of QPM on P-DukeMTMC under transfer setting. The results presented in Table IV show that QPM achieves state-of-the-art performance under all metrics. For example, QPM outperforms one of the most recent methods, i.e. PVPM [16], by 5.8% and 1.9% in terms of Rank-1 accuracy and mAP, respectively. Experiments on this database further justifies the effectiveness of QPM.

### D. Ablation Study

Ablation study are conducted on two large-scale datasets, i.e. Occluded-Duke [14] and P-DukeMTMC [20]. The experimental results on the reported two benchmarks are shown in Table V. It should be noted that Table V lists the results on Occluded-Duke under supervised setting and the results on P-DukeMTMC under transfer setting. These results show the robustness and effectiveness of our method under different experimental settings.



TABLE III  
PERFORMANCE COMPARISONS ON PARTIAL-iLIDS AND PARTIAL-REID UNDER TRANSFER SETTING. + REPRESENTS THE EXTENDED VERSION OF THE CONFERENCE PAPER.

Methods		Partial-iLIDS		Partial-REID	
		R-1	R-3	R-1	R-3
partial	MTRC [62]	17.7	26.1	23.7	27.3
	AMC+SWM [19]	21.0	32.8	37.3	46.0
	DSR [18]	58.8	67.2	50.7	70.0
	DCR [63]	60.5	69.7	52.0	67.5
	STNReID [64]	54.6	71.3	66.7	80.3
	SFR [28]	63.9	74.8	56.9	78.5
	VPM [31]	65.5	74.8	67.7	81.9
	HOReID [15]	72.6	86.4	85.3	91.0
	<b>QPM</b>	<b>77.3</b>	<b>85.7</b>	<b>81.7</b>	<b>88.0</b>
original	MaskReID [22]	33.0	-	28.7	-
	PCB [3]	46.8	-	56.3	-
	ISP* [30]	66.4	80.7	65.7	75.3
	PGFA [14]	69.1	80.9	68.0	80.0
	PVPM [16]	-	-	78.3	-
	PGFA+ [59]	70.6	81.3	72.5	83.0
	<b>QPM</b>	<b>77.3</b>	<b>85.7</b>	<b>81.7</b>	<b>88.0</b>

TABLE IV  
PERFORMANCE COMPARISONS ON P-DukeMTMC UNDER TRANSFER SETTING.

Methods	Rank-1	Rank-5	Rank-10	mAP
HACNN [2]	30.4	42.1	49.0	17.0
MLFN [65]	31.3	43.6	49.6	18.1
OsNet [66]	33.7	46.5	54.0	20.1
IDE [54]	36.0	49.3	55.2	19.7
Part Bilinear [56]	39.2	50.6	56.4	25.4
PCB [3]	43.6	57.1	63.3	24.7
PGFA [14]	44.2	56.7	63.0	23.1
ISP* [30]	46.3	56.9	60.8	26.4
PGFA+ [59]	48.2	59.6	65.8	26.8
PVPM [16]	51.5	64.4	69.6	29.2
<b>QPM</b>	<b>57.3</b>	<b>69.9</b>	<b>75.5</b>	<b>31.1</b>

$$\mathcal{L} = \mathcal{L}_{part}^{id} + \mathcal{L}_{part}^{tp} + \mathcal{L}_{global}^{id} + \mathcal{L}_{sg}^{id} + \mathcal{L}_{global}^{tp}. \quad (16)$$

**Effectiveness of the quality scores.** In Table V, ‘Baseline’ refers to the PCB model [3]. ‘Baseline(+triplet)’ equips PCB with the triplet loss in Eq. 3, while all part quality scores are set to 1. ‘Part branch’ means that we adopt the part feature learning branch only in QPM for ReID. As shown in Table V, ‘Baseline(+triplet)’ slightly improves the performance relative to the baseline. In comparison, the ‘Part branch’ brings in significant performance promotion, suggesting that part quality scores considerably benefit the occluded ReID task.

**Effectiveness of the ISA module.** In Table V, ‘GAP global’ means that we perform GAP on the feature maps  $\mathbf{G}$  to obtain the global feature for each image. ‘AGFE global’ means that we utilize adaptive global feature extraction module. When the ISA module is equipped, performance of both types of global features is promoted. In particular, ISA promotes the Rank-1 accuracy of ‘AGFE global’ by 6.0% and 4.6%, as well as mAP by 3.0% and 1.8%, on the two databases, respectively.

**Effectiveness of the AGFE module.** Compared with ‘GAP global’ and ‘GAP global(+ISA)’, the AGFE module consistently brings about significant performance gains. For example, ‘AGFE global(+ISA)’ outperforms ‘GAP global(+ISA)’ in the

TABLE V  
ABLATION STUDY ON EACH KEY COMPONENT OF QPM.

Method	Occluded-Duke		P-DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
Baseline	42.1	34.8	47.2	22.9
Baseline(+triplet)	43.7	37.0	48.5	26.4
Part branch	58.3	46.7	55.1	28.8
GAP global	40.8	31.7	26.0	14.7
GAP global(+ISA)	42.7	33.0	27.0	15.5
AGFE global	56.6	42.2	41.4	23.3
AGFE global(+ISA)	62.6	45.2	46.0	25.1
<b>QPM</b>	<b>64.4</b>	<b>49.7</b>	<b>57.3</b>	<b>31.1</b>

TABLE VI  
PERFORMANCE COMPARISON BETWEEN DIFFERENT TYPES OF GLOBAL FEATURES.

Method	Occluded-Duke		P-DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
GAP global	40.8	31.7	26.0	14.7
SI global	38.5	26.6	26.5	13.1
AGFE global	56.6	42.2	41.4	23.3

Rank-1 accuracy by as much as 19.9% on Occluded-Duke. These experimental results indicate that it is vital to adaptively extract global features from the common non-occluded regions for each image pair.

Moreover, we compare the performance of ‘AGFE global’ with ‘SI global’ in Table VI. ‘SI global’ means that we obtain global features for each single image (SI) using Eq. 4, without considering the difference in occlusion locations for each image pair. To facilitate a fair comparison, ‘SI global’ adopts the same loss functions as ‘AGFE global’, i.e., the cross-entropy loss and triplet loss. It is shown that the performance of ‘SI global’ drops dramatically compared with ‘AGFE global’, which suggests that Eq. 4 alone cannot promote ReID performance. This experimental result indicates that it is vital to extract semantically consistent global features for each image pair.

**Effectiveness of the combination.** With both ISA and AGFE modules, the quality of global features is promoted significantly. For example, ‘AGFE global(+ISA)’ outperforms ‘GAP global’ in the Rank-1 accuracy by as much as 21.8% and 20.0% on Occluded-Duke and P-DukeMTMC databases, respectively.

Finally, the combination of the part branch and the global branch, which is denoted as QPM in Table V, achieves better performance than using either one branch alone. The above comparisons justify the effectiveness of each key component in QPM.

**ISA vs. Other Approaches.** To facilitate fair comparison, all experiments are based on the ‘AGFE global’ model. We equip the ‘AGFE global’ model with different spatial attention modules respectively and summarize their performance in Table VII. It is shown that ISA significantly outperforms all other methods by at least 3.3% and 2.9% in terms of Rank-1 accuracy on Occluded-Duke and P-DukeMTMC, respectively. This is because ISA is identity-aware; therefore, it effectively handles the occlusions between pedestrians.



TABLE VII  
PERFORMANCE COMPARISONS OF DIFFERENT ATTENTION MODELS.

Methods	Occluded-Duke		P-DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
AGFE global	56.6	42.2	41.4	23.3
+FC-SA [67]	56.9	42.0	42.8	23.8
+RGA-SA [68]	57.6	41.2	40.9	23.1
+RA-SA [69]	58.1	42.7	41.3	23.3
+CBAM-SA [70]	59.3	43.9	43.1	24.5
<b>+ISA</b>	<b>62.6</b>	<b>45.2</b>	<b>46.0</b>	<b>25.1</b>

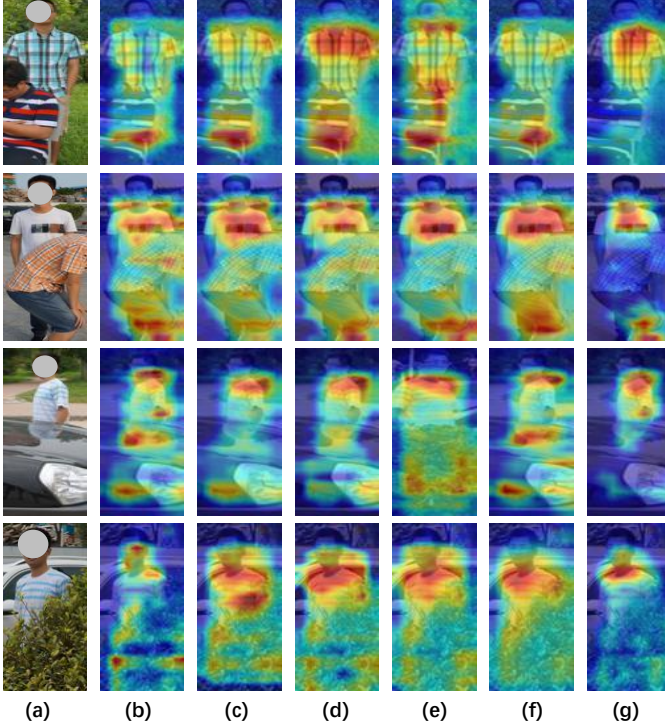


Fig. 5. Illustration of heat maps for feature maps after processing by different attention models. (a) The original image. (b) w/o attention. (c) FC-SA. (d) RGA-SA. (e) RA-SA. (f) CBAM-SA. (g) ISA (ours). (Best viewed in color.)

Moreover, we illustrate the heat maps for feature maps after the processing of different attention modules in Fig. 5. We have the following observations. First, heat maps for the baseline model have high responses on both occluded and non-occluded regions; second, existing popular spatial attention models [67]–[70] handle occlusions between pedestrians poorly; third, with the identity-aware guidance, our ISA module can well differentiate discriminative body parts from the occluded ones by both objects and other pedestrians. The above comparisons further demonstrate the effectiveness of ISA.

**Comparisons of Model Complexity.** In this experiment, we demonstrate that QPM not only achieves superior performance in terms of occluded ReID accuracy, but also offers advantages in terms of both its time and space complexities.

Three recent powerful occluded ReID approaches are compared: PGFA [14], HOREID [15], and PGFA<sup>+</sup> [59]. To facilitate fair comparison, all experimental settings are consistent with the paper description. Following [14], [15], [59], the input

TABLE VIII  
COMPARISONS OF MODEL COMPLEXITY ON THE OCCLUDED-DUKE DATABASE. + REPRESENTS THE EXTENDED VERSION OF THE CONFERENCE PAPER.

Method	Train	Inference	Params	Rank-1
PGFA [14]	0.82s/iter	0.78s/img	115.5M	51.4
HOREID [15]	1.34s/iter	1.94s/img	117.5M	55.1
PGFA <sup>+</sup> [59]	0.88s/iter	0.79s/img	115.5M	56.3
<b>QPM</b>	<b>0.68s/iter</b>	<b>0.47s/img</b>	<b>29.0M</b>	<b>64.4</b>

TABLE IX  
PERFORMANCE OF QPM WITH DIFFERENT VALUES OF  $K$ .

Part number $K$	Occluded-Duke		P-DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
2	61.7	49.1	44.2	25.6
3	61.9	49.4	49.1	26.8
4	62.7	49.4	50.5	28.4
<b>6</b>	<b>64.4</b>	<b>49.7</b>	<b>57.3</b>	<b>31.1</b>
8	62.7	49.5	53.7	28.8

images are resized to  $256 \times 128$  pixels for HOREID and  $384 \times 128$  pixels for PGFA, PGFA<sup>+</sup> and QPM. The batch size is set to 64 uniformly for all methods. Comparisons are conducted on a Titan V GPU, and results are summarized in Table VIII. The inference time cost in Table VIII includes the feature extraction of the query image and the matching time of all gallery images.

As is evident from the Table VIII, the model size of QPM is much smaller since it does not need an additional detection model. In addition, QPM has a faster training and testing speed. Specifically, our test time is only 60% and 24% of PGFA [14] and HOREID [15], respectively. This is because QPM does not require time-consuming human key point extraction. Although QPM uses less additional information during training and testing than PGFA [14] and HOREID [15], it still has significant performance advantages. Accordingly, the above comparisons demonstrate that the proposed QPM model is both compact and efficient.

#### E. Parameter Analysis

**The Impact of Part Number  $K$ .** In this experiment, we analyze the impact of the part number  $K$ . Experimental results are summarized in Table IX. It is shown that the optimal value of  $K$  is 6, which is consistent with the conclusion in [3]. Therefore, we consistently set  $K$  to 6 in this work.

**The Impact of the Feature Dimension  $d$ .** Table X shows the ReID performance with different feature dimension  $d$ . QPM

TABLE X  
PERFORMANCE OF QPM WITH DIFFERENT VALUES OF  $d$ .

Feature dimension $d$	Occluded-Duke		P-DukeMTMC	
	rank-1	mAP	rank-1	mAP
128	61.0	48.7	52.0	28.8
256	62.9	49.5	55.2	29.3
512	63.5	49.7	56.3	30.8
<b>1024</b>	<b>64.4</b>	<b>49.7</b>	<b>57.3</b>	<b>31.1</b>

TABLE XI  
PERFORMANCE OF QPM WITH DIFFERENT VALUES OF  $\gamma$ .

Weight $\gamma$	Occluded-Duke		P-DukeMTMC	
	rank-1	mAP	rank-1	mAP
0.5	64.0	49.3	57.0	30.7
<b>0.6</b>	<b>64.4</b>	<b>49.7</b>	<b>57.3</b>	<b>31.1</b>
0.7	63.6	49.3	57.1	30.6
0.8	62.9	49.3	56.6	30.1
0.9	61.4	48.7	56.2	29.6
1.0	58.3	46.7	55.1	28.8

consistently achieves state-of-the-art performance when  $d$  is set to 256, 512, and 1024, with the best result achieved when  $d$  is set to 1024.

**The Impact of the weight  $\gamma$ .** Table XI shows the ReID performance with different values of  $\gamma$ . It is shown that the optimal value of  $\gamma$  is 0.6. Therefore, we consistently set  $\gamma$  to 0.6 in this work.

## V. CONCLUSION

In this paper, we propose a novel framework named QPM to handle the occluded person ReID problem. Unlike most existing methods, which depend on visibility cues from outside tools, QPM jointly learns part features and predicts part quality in an end-to-end framework without using any annotations or outside tools. Moreover, based on the predicted part quality scores, we propose a novel identity-aware spatial attention (ISA) model to handle occlusion between pedestrians. We further design a novel approach that adaptively generates global features from common non-occluded regions for each image pair. Finally, extensive experiments on four popular datasets demonstrate the effectiveness of QPM.

## REFERENCES

- [1] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 667–676.
- [2] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2285–2294.
- [3] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.
- [4] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2119–2128.
- [5] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2590–2600.
- [6] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1179–1188.
- [7] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5794–5803.
- [8] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2588–2603, 2017.
- [9] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.
- [10] W. J. Scheirer, P. J. Flynn, C. Ding, G. Guo, V. Struc, M. Al Jazaery, K. Grm, S. Dobrsek, D. Tao, Y. Zhu *et al.*, "Report on the btas 2016 video person recognition evaluation," in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016.
- [11] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, 2018.
- [12] C. Yan, G. Pang, X. Bai, C. Liu, N. Xin, L. Gu, and J. Zhou, "Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss," *IEEE Trans. Multimedia*, 2021.
- [13] B. Jiang, X. Wang, A. Zheng, J. Tang, and B. Luo, "Ph-gcn: Person retrieval with part-based hierarchical graph convolutional network," *IEEE Trans. Multimedia*, 2021.
- [14] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 542–551.
- [15] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020.
- [16] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person reid," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020.
- [17] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.
- [18] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7073–7082.
- [19] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4678–4686.
- [20] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *2018 IEEE Int. Conf. Multimedia Expo*, 2018, pp. 1–6.
- [21] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1062–1071.
- [22] L. Qi, J. Huo, L. Wang, Y. Shi, and Y. Gao, "Maskreid: A mask based deep ranking neural network for person re-identification," *arXiv preprint arXiv:1804.03864*, 2018.
- [23] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, 2019.
- [24] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4099–4108.
- [25] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3960–3969.
- [26] L. He, Y. Wang, W. Liu, X. Liao, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019.
- [27] L. He and W. Liu, "Guided saliency feature learning for person re-identification in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 357–373.
- [28] L. He, Z. Sun, Y. Zhu, and Y. Wang, "Recognizing partial biometric patterns," *arXiv preprint arXiv:1810.07399*, 2018.
- [29] J. Zhuo, J. Lai, and P. Chen, "A novel teacher-student learning framework for occluded person re-identification," *arXiv preprint arXiv:1907.03253*, 2019.
- [30] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 346–363.
- [31] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 393–402.
- [32] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 274–282.
- [33] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8514–8522.
- [34] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 420–429.
- [35] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided

- feature decomposition and fusion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1077–1085.
- [36] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3219–3228.
- [37] D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning deep context-aware features over body and latent parts for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 384–393.
- [38] X. Gong, Z. Yao, X. Li, Y. Fan, B. Luo, J. Fan, and B. Lao, “Lag-net: Multi-granularity network for person re-identification via local attention system,” *IEEE Trans. Multimedia*, 2021.
- [39] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, “Diverse part discovery: Occluded person re-identification with part-aware transformer,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2898–2907.
- [40] C. Wan, Y. Wu, X. Tian, J. Huang, and X.-S. Hua, “Concentrated local part discovery with fine-grained part representation for person re-identification,” *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1605–1618, 2019.
- [41] C. Ding, K. Wang, P. Wang, and D. Tao, “Multi-task learning with coarse priors for robust part-aware person re-identification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
- [42] K. Wang, P. Wang, C. Ding, and D. Tao, “Batch coherence-driven network for part-aware person re-identification,” *IEEE Trans. Image Process.*, vol. 30, pp. 3405–3418, 2021.
- [43] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
- [44] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, “Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10407–10416.
- [45] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, “See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4747–4756.
- [46] Y. Liu, J. Yan, and W. Ouyang, “Quality aware network for set to set recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5790–5799.
- [47] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [48] S. Zhao, C. Gao, J. Zhang, H. Cheng, C. Han, X. Jiang, X. Guo, W.-S. Zheng, N. Sang, and X. Sun, “Do not disturb me: Person re-identification under the interference of other pedestrians,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 647–663.
- [49] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, “See more, know more: Unsupervised video object segmentation with co-attention siamese networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3623–3632.
- [50] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, “Information fusion for classification of hyperspectral and lidar data using ip-cnn,” *IEEE Trans. Geosci. Remote. Sens. Lett.*, 2021.
- [51] P. Xie, M. Zhao, and X. Hu, “Pisltrc: Position-informed sign language transformer with content-aware convolution,” *IEEE Trans. Multimedia*, 2021.
- [52] W.-S. Zheng, S. Gong, and T. Xiang, “Person re-identification by probabilistic relative distance comparison,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 649–656.
- [53] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3754–3762.
- [54] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, “Person re-identification in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1367–1376.
- [55] K. Zhou and T. Xiang, “Torchreid: A library for deep learning person re-identification in pytorch,” *arXiv preprint arXiv:1910.10093*, 2019.
- [56] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, “Part-aligned bilinear representations for person re-identification,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–437.
- [57] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13 001–13 008.
- [58] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, “Adversarially occluded samples for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5098–5107.
- [59] J. Miao, Y. Wu, and Y. Yang, “Identifying visible parts via pose estimation for occluded person re-identification,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.
- [60] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
- [61] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.
- [62] S. Liao, A. K. Jain, and S. Z. Li, “Partial face recognition: Alignment-free approach,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, 2012.
- [63] Z. Gao, H. Zhang, L. Gao, Z. Cheng, R. Hong, and S. Chen, “Dcr: A unified framework for holistic/partial person reid,” *IEEE Trans. Multimedia*, 2020.
- [64] H. Luo, W. Jiang, X. Fan, and C. Zhang, “Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification,” *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2905–2913, 2020.
- [65] X. Chang, T. M. Hospedales, and T. Xiang, “Multi-level factorisation net for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2109–2118.
- [66] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, “Omni-scale feature learning for person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3702–3712.
- [67] Y. Liu, Z. Yuan, W. Zhou, and H. Li, “Spatial and temporal mutual promotion for video-based person re-identification,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 8786–8793.
- [68] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, “Relation-aware global attention for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3186–3195.
- [69] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3156–3164.
- [70] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “Cbam: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.