Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification

Jiaxu Miao, Yu Wu, and Yi Yang

Abstract—**We focus on the occlusion problem in person re-identification (re-id), which is one of the main challenges in real-world person retrieval scenarios. Previous methods on the occluded re-id problem usually assume that only the probes are occluded, thereby removing occlusions by manually cropping. However, this may not always hold in practice. This paper relaxes this assumption and investigates a more general occlusion problem, where both the probe and gallery images could be occluded. The key to this challenging problem is depressing the noise information by identifying bodies and occlusions. We propose to incorporate the pose information into the re-id framework, which benefits the model in three aspects. First, it provides the location of the body. We then design a *Pose-Masked Feature Branch* to make our model focus on the body region only and filter those noise features brought by occlusions. Second, the estimated pose reveals which body parts are visible, giving us a hint to construct more informative person features. We propose a *Pose-Embedded Feature Branch* to adaptively re-calibrate channel-wise feature responses based on the visible body parts. Third, in testing, the estimated pose indicates which regions are informative and reliable for both probe and gallery images. Then we explicitly split the extracted spatial feature into parts. Only part features from those *commonly visible parts* are utilized in the retrieval. To better evaluate the performances of the occluded re-id, we also propose a large-scale dataset for the occluded re-id with more than 35,000 images, namely Occluded-DukeMTMC. Extensive experiments show our approach surpasses previous methods on the occluded, partial, and non-occluded re-id datasets.**

Index Terms—**Occluded Person Re-Identification, human pose, occlusion detection.**

I. Introduction

PERSON re-identification (re-id) is a popular computer vision task, which aims at searching people across non-overlapping camera views at different times. Although recent approaches have achieved great progress, person re-id still suffers from large varieties of occlusions, pose, illumination, and so on. Occlusion is one of the main challenges for the person re-id since the occlusions introduce distractive information and confuse the re-id models.

Most existing person re-id methods [1], [2], [3], [4] use features of the whole pedestrian images for retrieval. However, the distractive occlusion information may also be encoded in these global features. Thus these models are not robust when meeting the occlusion situations. For instance, in Fig. 1, if a probe person image is occluded by a tree, the methods that cannot distinguish the target person and obstacles will retrieve incorrect results with a similar tree.

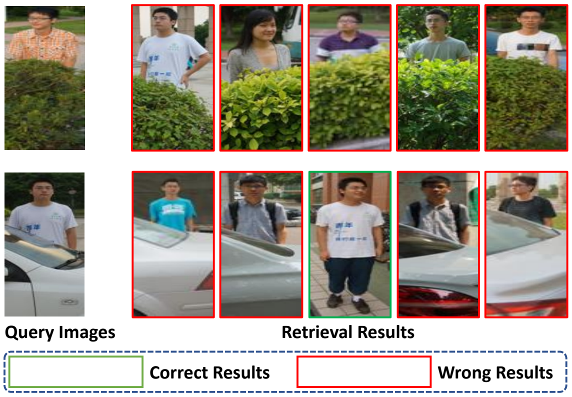

Fig. 1. Failure cases of previous methods [2] when meeting the occlusion situation.

Some partial re-id methods [5], [6], [7], [8] are proposed to tackle the occlusion problem. The partial re-id assumes that only query images contain occlusions, while all the gallery images are non-occluded. The occluded query images are manually cropped, and the visible part remains as the new query images. Thus the partial re-id aims at searching the same person in full-body appearance given only a partial probe image. This is called the *partial person re-id* problem. Although these partial re-id methods move a significant step towards solving the occlusion problem, there are some limitations: (1) The assumption that only probe images are occluded is too strong and not always hold in practice. (2) They need manually cropping, which is time-consuming, especially there are a large number of occlusion images. Although some very recent works [9], [10] do not require manually cropping, however, they still assume that all the gallery images are non-occluded.

In this paper, we relax the assumption and suppose that both probes and gallery images contain occlusions, namely the *occluded person re-id* problem. Fig. 2 shows the comparison between the partial and occluded re-id problem. Differently, the gallery set in our setting contains both occluded images and non-occluded images, which is in accordance with the practical application. Our setting is more challenging since there are difficult cases that both probe and gallery are occluded. Under the new assumption of the occluded re-id, we construct a large-scale dataset, **Occluded-DukeMTMC**, in which the gallery set consists of both occluded and non-occluded images while all probe images are occluded.

To tackle the challenging occluded person re-id problem, we design a learnable person retrieval system by Convolutional Neural Networks (CNN). A fundamental solution for the occluded re-id is to depress the noise information brought by occlusions. Thus, three specially designed NN modules

Jiaxu Miao, Yu Wu and Yi Yang are with ReLER, the Australian Artificial Intelligence Institute, University of Technology Sydney. (email: jiaxu.miao@student.uts.edu.au; yu.wu-3@student.uts.edu.au; yi.yang@uts.edu.au.)

for the occluded re-id are proposed, utilizing CNN features' properties in the spatial and channel dimension. Spatially, for CNN features, the information of the target person and occlusions is located in the corresponding locations [11]. Thus an attention mask can help filter out the information of occlusions. Channel-wisely, channels of the CNN features contain different information [12] and re-calibration of the channel feature can depress the occlusion information.

Concretely, we utilize a pre-trained human pose estimation model to provide key-point landmarks, which are used to improve the re-id framework in the following three aspects.

First, the pose estimation provides the location of the body. If a part of a person is occluded, the corresponding landmark is missing (with low confidence score) in the pose estimation results. We then generate spatial masks based on the visible landmarks and then filter the CNN feature maps by these masks. Therefore, we can filter the noise information brought by occlusions and makes our model focus on the body region only. This is the *Pose-Masked Feature* Branch.

Second, estimated visible pose landmarks help to construct more informative person features. Based on the information that which body part is visible and which one is occluded, we propose the *Pose-Embedded Feature* Branch to enhance the learned features by dynamically adjusting the response of the CNN channels. Specifically, we first generate a pose embedding by the visible landmarks. Then the embedding is utilized as gates to adaptively re-calibrate channel-wise feature responses. By the gating operation, the related channels of the visible parts are activated, while those channels of missed parts are further depressed.

Third, in testing, the estimated pose indicates which regions are informative and reliable for both probe and gallery images. Then we explicitly split the extracted spatial feature into parts. Only part features from those *commonly visible parts* are utilized to calculate the distance in the retrieval. In this way, we further reduce the impact of the occlusions.

In our conference version [13], we only focus on the spatial information provided by the pose model, *i.e.*, the Pose-Masked Feature Branch and the matching on commonly visible parts. However, these two modules are mostly based on hard spatial operations. Although depressing the spatial feature where the occlusions locate alleviates the occlusion problem in person re-id, the conference version [13] ignores the critical information of the feature channels. Motivated by SENet [12], in this paper, we improve the previous method by introducing the Pose-Embedded feature branch, which enhances features by adaptively adjusting channel-wise feature responses based on the visible parts. Concretely, we employ visible landmarks to generate the channel weights, which are used to re-calibrate the channel-wise feature responses and enhance the non-occluded information. The re-calibration operation on the channel dimension selectively emphasizes informative features and suppresses less useful ones generated by occlusions. Extensive ablation experiments show the effectiveness of this new branch.

We conduct extensive experiments on five person re-id datasets, including occluded, partial, and non-occluded datasets. Results show that our method not only surpasses previous re-id approaches [5], [6], [7], [8] on occluded and partial datasets, but also shows the superiority on two non-occluded datasets. Our contributions are summarized as follows:

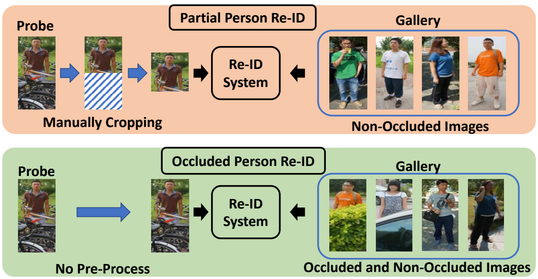

Fig. 2. Comparison of the partial (above) and occluded re-id problem (below).

- We propose a large-scale occluded re-id dataset, Occluded-DukeMTMC, which is facilitative for studies on the occluded re-id problem.
- We design a learnable person retrieval system for the occluded re-id, utilizing the properties of NN features in the spatial and channel dimension.
- We propose the Pose-Masked Feature Branch, which spatially removes the noise information brought by occlusions.
- We propose the Pose-Embedded Feature Branch, which generates better re-id features by adaptively adjusting channel-wise feature responses based on the visible parts.
- We propose only to consider the commonly visible parts when calculating the distance for retrieval.

II. Related Work

Deep Person Re-ID. With the development of deep learning [14], many deep learning-based person re-id approaches [15], [16], [17], [18], [19], [20], [2], [3], [21], [1], [22], [23], [24], [25], [26] have been proposed and show significant superiority on retrieval accuracy. Some deep re-id approaches use one global feature learned by the classification loss [27], [28] or enhanced by the triplet loss [29] or the quadruplet loss [30]. Recently, some part-based person re-id approaches [2], [31], [32], [33], [34], [35], [36] have been proposed. For instance, Zhao *et al.* [34] and Liu *et al.* [35] employ an attention mechanism to extract partial features. Kalayeh *et al.* [33] propose to utilize human parsing methods and extract the feature for each human part. The final representation is assembled by human-part features. Sun *et al.* [2] propose PCB, which horizontally split the feature map for constructing partial features. Wang *et al.* [32] and Fu *et al.* [31] partition multiple granularities of part features and further improve the retrieval performance. However, when dealing with the occluded re-id problem, these methods introduce the distractive occlusion information and may fail when occlusion occurs.

Partial Person Re-ID. Partial person re-id approaches [5], [6], [7], [8] have been proposed for solving the occlusion problem in person re-id. They assume that only probes contain occlusions, while gallery images are all non-occluded. The probes are manually cropped, and visible parts remain as new probes. Thus, the partial person re-id aims at searching a person with a partial image across non-occluded images.

For instance, Zheng *et al.* [5] firstly define the partial re-id problem. They propose local patch and global part matching to tackle this problem. He *et al.* [7], [8] introduce DSR and SFR, which use spatial feature reconstruction without time-consuming feature alignment. Sun *et al.* [6] propose visibility-aware part-level features for partial person re-id. This model is based on PCB and partitions the feature map horizontally. Then a region locater is learned to predict which partial feature is visible. Part of our method is similar to this paper. However, our method aims at solving the occluded re-id problem, and all the occlusion images are not pre-processed by manually cropping. Thus we use pose landmarks to indicate the occlusion region while Sun *et al.* [6] use a learnable region locater.

Occluded Person Re-ID. Recently some occluded person re-id approaches [9], [10] have been proposed. These methods have no manually cropping process and take as input the occluded person images directly. Zhuo *et al.* [9] propose to randomly simulate occlusion images in the training stage and use a classifier to predict if the input image is simulated or not. He *et al.* [10] propose Foreground-aware Pyramid Reconstruction for the occlusion re-id. This is an alignment-free method and achieves good performance. Although these methods move a significant step towards solving the occluded person re-id problem, they still assume that only probes are occluded. There is no very hard case that one occluded image is retrieved from occluded images. Our method relax this assumption and propose a corresponding dataset for the occluded person re-id.

Pose-Guided Person Re-ID. Pose landmarks are important information in person images and benefit the person re-id tasks. Many pose-guided methods [37], [38], [39], [40], [41] for person re-id have been proposed to facilitate person re-id models. These aforementioned pose-guided methods aim at tackling the human pose variation in the person re-id. Differently, our method employs landmarks for the occlusion situation in re-id. One method *et al.* [42] uses the attention mechanism [43] by pose information for the occlusion problem in the detection task—differently, our method focuses on the person re-id problem.

Object Tracking. Person re-id methods are also applicable to the object tracking task [44], [45], [46], and vice versa. For instance, The triplet loss [29] and quadruplet loss [30] can be transferred to the tracking task [45], [46]. The occlusion problem is also a critical problem on the tracking task, and several approaches [47], [48], [49] for the occlusion problem have been proposed. In this paper, we propose to use the pose landmarks to depress the occlusion information, which is applicable to the tracking task. However, since extracting pose landmarks is time-consuming, efficiency should be considered in the tracking task.

III. The Occluded-DukeMTMC Dataset

We propose a large-scale occluded re-id dataset where both the query and gallery images contain occlusions. The new dataset is derived from DukeMTMC-reID [50], [51].

Properties of Occluded-DukeMTMC. Most previous datasets [5], [52], [9] for the occlusion problem in person re-id are under the assumption that only probes are occluded while gallery images are non-occluded. This paper focuses on a more general occlusion problem in that both probes and gallery images contain occlusions. The gallery images contain both occlusion images and non-occluded images, according to the practical application. All the probe images are occluded, following the previous setting [5], [52], [9]. Thus, there exist hard cases that an occluded person image is compared with another occluded person image. Our setting is more difficult and practical compared with previous occlusion datasets [5], [52], [9]. Table. I shows the comparison between our Occluded-DukeMTMC and the previous occlusion datasets [5], [52]. Our constructed Occluded-DukeMTMC is the largest occluded re-id dataset to date. Previous datasets for partial re-id [5], [52] contains only hundreds of person images. Recently, Zhuo *et al.* [9] propose the Occluded-REID dataset for the occluded re-id, containing 2000 person images. Our Occluded-DukeMTMC contains $35,489$ person images, including $15,618$ images of 702 identities in the train set, $17,661$ images of $1,110$ identities in the gallery set and $2,210$ images of $1,110$ identities in the query set.

TABLE I
Comparison of three datasets on the occluded re-id problem.

Dataset	Train Set		Gallery Set		Query Set	
	Identity	Image	Identity	Image	Identity	Image
Partial-REID [5]	-	-	60	300	60	300
Partial-iLIDS [52]	-	-	119	119	119	119
Occ-DukeMTMC	702	15,618	1,110	17,661	519	2,210

Data Collection. Our Occluded-DukeMTMC is manually selected from DukeMTMC-reID [50], [51]. In the DukeMTMC-reID dataset, the occluded person image is less than 15% in the query set. Thus, it is not applicable to evaluate the occluded re-id. We manually select all the occluded person images from the query set and the gallery set in the original DukeMTMC-reID to construct the new query set. For the gallery set of our Occluded-DukeMTMC, we directly use the gallery set of the original DukeMTMC-reID, which contains 10% occluded images. Therefore, there exist the same images in query and gallery images. However, when evaluating the re-id approaches, the images with the same camera are ignored. Thus, there is no worry about retrieving the same image in the gallery. The train set in the Occluded-reID is selected from the train set of the original DukeMTMC-reID. In the train set of DukeMTMC-reID, there are some images containing exactly the same occlusions in the test set. These images may make the re-id model "remember" these specific occlusions and influence the generalization of the occluded re-id models. Thus, we manually remove all these 934 images from the DukeMTMC-reID dataset to construct the train set of our Occluded-DukeMTMC.

IV. Methodology

This paper address the occluded person re-id problem. To figure out which part of the person image is occluded, we employ the pose landmarks to identify the visible parts. When we extract pose landmarks, the landmarks in the occlusion region have a lower confidence score. Thus, we can obtain the visible landmarks which contain the occlusion information.

We strengthen the occluded re-id by using the visible landmarks in three aspects. First, the visible landmarks provide the location of visible parts and are used to generate spatial masks, which filter the noise introduced from occlusions. This is the Pose-Masked Feature Branch. Second, the visible landmarks contain the informative knowledge of occlusions, indicating which part is occluded while which part is visible. Thus, the visible landmarks are used to generate a pose embedding. The pose embedding is used as gates to adaptively re-calibrate channel-wise feature responses, selecting the visible channel features and depressing the invisible ones. This is called the Pose-Embedded Feature Branch. Third, in testing, the visible landmarks indicate which region is occluded. The global feature map is split into parts, and the commonly visible part features are utilized for comparison between the query and gallery images.

A. Preliminaries

The pipeline of our method is shown in Fig. 3. The backbone architecture of our method is ResNet50 [53], which removes the last average pooling layer and fully connected layers, following previous re-id approaches [2], [6]. Taking a person image I with a size of $H \times W$ as input, the original ResNet50 outputs a feature map with a spatial dimension of $H/32 \times W/32$. For extracting a more informative feature map, we enlarge the spatial size of the extracted feature map to $H/16 \times W/16$, by changing the stride of conv4_1 to 1 [2], [31]. A larger spatial feature map makes it easier to split the target person from the occlusions. The extracted feature map is denoted as $\mathbf{F}$.

B. Visible Landmarks Detection

As shown in Fig. 3, given a person image I, the pose landmarks are extracted by a human pose estimator, which is pre-trained on the COCO dataset [54]. Denote the number of the extracted pose landmarks as N, where $N = 18$ in this paper. The output of the pose estimator is the coordinates and confidence score of each landmark. When a landmark is occluded in the person image, the confidence score of the landmark is low. Thus, by setting a threshold γ, we can filter out the invisible landmarks and obtain visible human landmarks. [1] The visible landmarks contain the informative knowledge of the occlusions in the person image.

We utilize the location information of the visible landmarks for generating spatial masks in the Pose-Masked Feature Branch. Formally, the locations of the pose landmarks are

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

where S_j^{conf} and γ denote the confidence score and the threshold, respectively. $\mathbf{P}_j$ denotes the jth landmark location and cx_j,cy_j denote the coordinate of the jth landmark, $j = 1, ..., N$. Thus, we obtain visible landmarks $\mathbf{P}$ with the spatial location information.

For generating the pose embedding that contains the occlusion information in the Pose-Embedded Feature Branch, a visible landmark vector $\mathbf{p} \in \{0, 1\}^N$ is generated,

$$\mathbf{p}_j = \begin{cases} 1 & \text{if } S_j^{conf} \geq \gamma \\ 0 & \text{else,} \end{cases} \tag{2}$$

where S_j^{conf} and γ denote the confidence score and the threshold, respectively. Each element $\mathbf{p}_j$ of the visible landmark vector $\mathbf{p}$ denotes if the jth landmark is occluded or not. Thus, the visible landmark vector $\mathbf{p}$ is an encoding of the occlusion. We use $\mathbf{p}$ to generate the pose embedding, which encodes the occlusion information into the representation features to benefit the occluded re-id problem.

C. Pose-Masked Feature Branch

As shown in Fig. 3, the final pose-guided feature is obtained by a concatenation of three components, including the pose-masked feature in the Pose-Masked Feature Branch, the pose-embedded feature in the Pose-Embedded Feature Branch, and the global max-pooling feature of the feature map $\mathbf{F}$.

In the Pose-Masked Feature Branch, we use the positions of visible landmarks $\mathbf{P}$ to generate spatial masks. For the visible pose landmarks, which means $\mathbf{P}_j = (cx_j, cy_j)$, the generated pose mask is a Gaussian heatmap with the center at (cx_j, cy_j). For the invisible landmarks, where $\mathbf{P}_j = 0$, the spatial pose mask is set to $\mathbf{0}$. Denote each pose mask as $\mathbf{M}_j$, $j = 1, ..., N$. The pose masks $\mathbf{M}$ are downsampled to the spatial size of $\mathbf{F}$ by bi-linear interpolation. The feature map $\mathbf{F}$ multiply each pose mask $\mathbf{M}_j$ to generate N pose-masked feature maps $\mathbf{M}'_j$, $j = 1, ..., N$. The pose-masked feature maps filter out the occlusion parts and focus on the visible body parts, which depresses the information of the occlusions.

The pose-masked feature maps are fed into a max-pooling layer to generate N pose-masked feature vectors. Then N pose-masked feature vectors are max-pooled to generate one pose-masked feature $\mathbf{f}_{pm}$, as shown in Fig. 3. We utilize max-pooling instead of average pooling because the max-pooling operation ignores the occluded parts and redundant visible body parts indicated by pose landmarks.

D. Pose-Embedded Feature Branch

In the Pose-Embedded Feature Branch, we use the visible landmark vector $\mathbf{p}$ to generate the pose embedding. The pose embedding is used as channel gates for the global feature. Since the dimension of $\mathbf{p}$ is small (18) while the dimension of the global feature map $\mathbf{F}$ is large $(2,048)$, it is hard to encode the pose embedding properly. Thus we utilize a 1×1 convolutional layer to reduce the channel dimension of $\mathbf{F}$, and generate a new feature map $\mathbf{F}'$ with the channel dimension of $1,024$. We take as input $\mathbf{p}$ and use an embedding encoder network with two fully-connected layers and a sigmoid activation layer in the end to generate the pose embedding $\mathbf{p}'$. The pose embedding $\mathbf{p}'$ is used as channel gates and multiplies $\mathbf{F}'$ channel-wisely to generate the pose-embedded feature map

[1] When a person is occluded by another one, we choose the person with a larger number of visible landmarks as the target person.

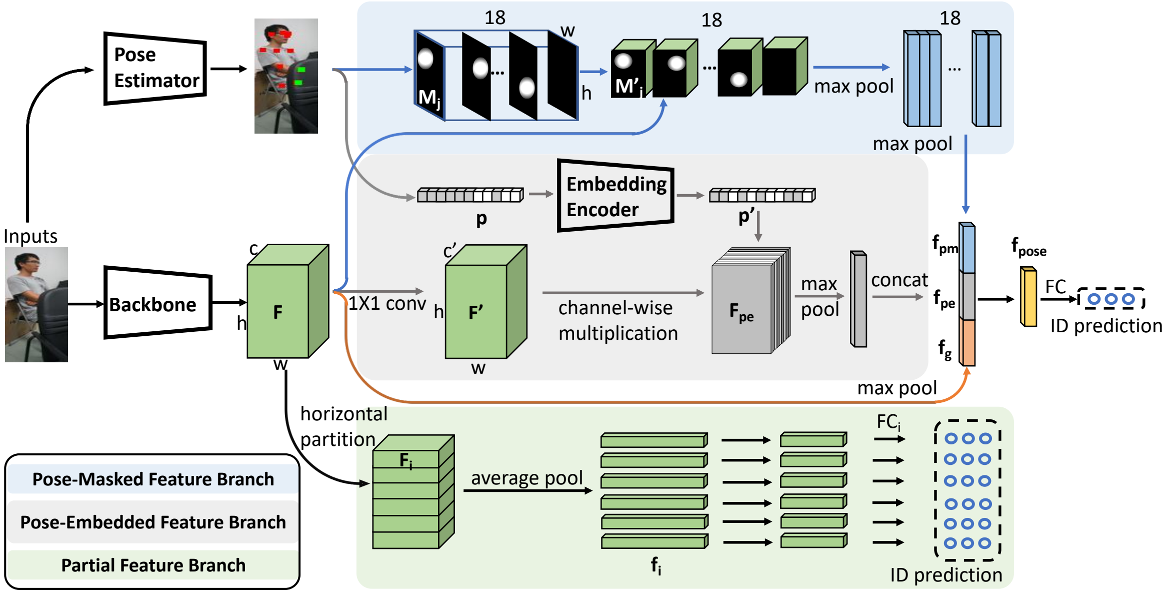

Fig. 3. The pipeline of the proposed method. Red and green points indicate *visible* and *invisible* landmarks, respectively. Our model contains three branches. In Pose-Masked Feature Branch, we generate Gaussian maps to filter out occlusions. In the Pose-Embedded Feature Branch, we obtain the pose-embedding by the visible landmark vector. These embeddings are further used to generate channel gates, which control the response of channels by the channel-wise multiplication. In Partial Feature Branch, we uniformly split the extracted feature map into parts for generating part features.

$\mathbf{F}_{pe}$. A max-pooling layer is employed to generate the pose-embedded feature $\mathbf{f}_{pe}$. The pose-embedded feature $\mathbf{f}_{pe}$ encodes the occlusion information implicitly.

Occlusion Simulation. In our setting, the query and gallery sets contain plenty of occlusion images. However, the train set contains a few occlusions, resulting in insufficient varieties of the visible landmark vectors. Thus, we simulate the occluded person images in the train set and generate corresponding visible landmark vectors $\mathbf{p}$, which improves our model to learn the knowledge of occlusions. The occluded person images are simulated by randomly erasing the person images. The landmarks in the erased areas are annotated as invisible landmarks.

E. Optimization

The pose-guided feature is obtained by a concatenation of three components, including the pose-masked feature $\mathbf{f}_{pe}$ in the Pose-Masked Feature Branch, the pose-embedded feature $\mathbf{f}_{pm}$ in the Pose-Embedded Feature Branch, and the global max-pooling feature $\mathbf{f}_g$ of the feature map $\mathbf{F}$.

We reduce the dimension of the concatenated feature to 256 by a fully connected layer and obtain the pose-guided feature $\mathbf{f}_{pose}$. A fully connected layer and a softmax layer are used to predict the identity of the person image. Denote the prediction of these two pose-guided branches as $\hat{y}$. The loss function of these pose-guided branches are

$$\mathcal{L}_{pose} = CE(\hat{y}, y) \,, \tag{3}$$

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

Except for the Pose-Masked Feature Branch and the Pose-Embedded Feature Branch, we also use a Partial Feature Branch to obtain the discriminative part features, as shown in Fig. 3. In the Partial Feature Branch, the partial feature maps $\mathbf{F}_i$ are obtained by splitting the feature map $\mathbf{F}$ into p parts, $i = 1, ..., p$. Then the partial feature map $\mathbf{F}_i$ is fed into an average pooling layer to generate the partial feature vector $\mathbf{f}_i$. We reduce the dimension of the partial feature vector $\mathbf{f}_i$ to 256. We use a fully connected layer and a softmax layer to predict the identity of the person image. Thus, we can obtain the objective function for the Partial Feature Branch $\mathcal{L}_{part}$ by

$$\mathcal{L}_{part} = \sum_{i=1}^{p} CE(\hat{y}_i, y), \tag{4}$$

where $\hat{y}_i$ is the prediction based on the i-th part. The total objective function $\mathcal{L}$ is a linear combination of the two losses,

$$\mathcal{L} = \lambda\mathcal{L}_{part} + (1-\lambda)\mathcal{L}_{pose} \,, \tag{5}$$

where λ denotes a coefficient to balance $\mathcal{L}_{part}$ and $\mathcal{L}_{pose}$.

F. Feature Matching in Commonly Visible Parts

Fig. 4 shows the matching strategy in testing. In Fig. 4, the visible pose landmarks indicate the visible partial features. The part containing at least one visible landmark is annotated as a visible part. Thus, we can compare the probe and gallery images using partial features in the commonly visible parts. This operation filters out the distractive information of occlusions. Besides, we also use the constructed pose-guided feature for distance computation. We average the distances calculated by visible partial features and the pose-guided feature.

Formally, for a partial feature $\mathbf{f}_i$, $i = 1, ..., p$, we can obtain a visible label $l_i \in \{0, 1\}$ which represents if this part is occluded or not. For each body part i,

$$l_i = \begin{cases} 1 & \text{if } \exists cy_j \in [\frac{i-1}{p}H, \frac{i}{p}H) \\ 0 & \text{else} \end{cases} \quad (j = 1, ..., N) \,, \tag{6}$$

where cy_j is the jth longitudinal coordinate of landmark $\mathbf{P}_j$ and H is the height of the image. The distance of the ith part between the query and gallery is,

$$d_i = D(\mathbf{f}_i^p, \mathbf{f}_i^g) \quad (i = 1, ..., p) \,, \tag{7}$$

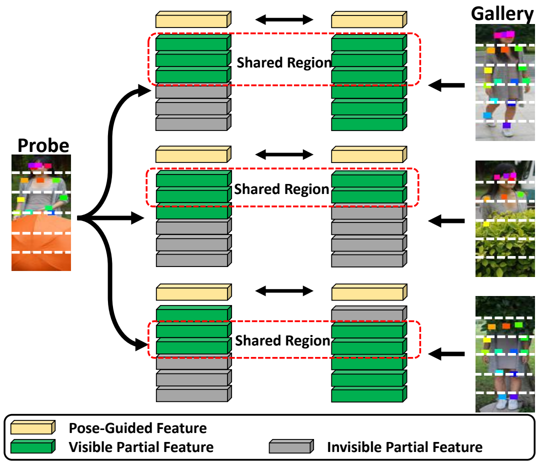

Fig. 4. Matching strategy of our method. The partial features in the commonly visible parts and the pose-guided feature are used for distance computation.

where $D(\cdot)$ is the cosine distance function, $\mathbf{f}_i^p$, $\mathbf{f}_i^g$ are the ith partial feature of the probe and gallery images, respectively.

Besides, we compute the pose-guided distance by,

$$d_{pose} = D(\mathbf{f}_{pose}^p, \mathbf{f}_{pose}^g)\,, \tag{8}$$

where $\mathbf{f}_{pose}^p$, $\mathbf{f}_{pose}^g$ are the pose-guided feature of the probe and gallery images, respectively.

The total distance is computed by averaging the distances of the partial feature in the commonly visible parts and the pose-guided feature.

$$dist = \frac{\sum_{i=1}^{p}(l_i^p \cdot l_i^g)d_i + d_{pose}}{\sum_{i=1}^{p} l_i^p \cdot l_i^g + 1}\,, \tag{9}$$

where $dist$ denotes the final distance, l_i^p and l_i^g are the ith visible indicator of the probe and gallery images, respectively.

V. Experiments

A. Datasets and evaluation metrics

We evaluate our method on five datasets, including occluded, partial, and non-occluded re-id datasets.

Occluded-DukeMTMC is our proposed occluded re-id dataset, which consists of $15,618$ images in the train set, $17,661$ images in the gallery set, and $2,210$ images in the query set. Evaluation on this dataset illustrates the effectiveness of our approach on the occluded re-id task. **Partial-REID** [5] is a partial re-id dataset, with 600 person images of 60 identities. 300 person images are occluded, while 300 person images are non-occluded in Partial-REID. **Partial-iLIDS** [52] is a simulated partial re-id dataset including 238 person images of 119 person identities. Each ID contains an occluded person image and a non-occluded person image, respectively. For Partial-REID and Partial-iLIDS, in partial re-id approaches, the occluded images are cropped, and the visible parts are collected as new query images. In our setting, the cropping process is unnecessary. **Market-1501** [55] consists of $32,668$ images of $1,501$ person identities. Most of these images are non-occluded. Thus, Market-1501 is a non-occluded re-id dataset. **DukeMTMC-reID** [50], [51] consists of $36,411$ person images of $1,404$ identities. There exists some occluded images in DukeMTMC-reID. However, the query set of DukeMTMC-reID only contains less than 15% occluded images. Thus, DukeMTMC-reID can be viewed as a non-occluded dataset.

TABLE II
Results on Occluded-DukeMTMC.

Method	Rank-1	Rank-5	Rank-10	mAP
DIM [56]	21.5	36.1	42.8	14.4
LOMO+XQDA [57]	8.1	17.0	22.0	5.0
Part Aligned [34]	28.8	44.6	51.0	20.2
Random Erasing [58]	40.5	59.6	66.8	30.0
HACNN [59]	34.4	51.9	59.4	26.0
Triplet [29]	35.5	52.8	61.1	27.0
Aligned reID [60]	41.5	58.8	65.7	32.7
Adver Occluded [61]	44.5	-	-	32.2
PCB [2]	42.6	57.1	62.9	33.7
Part Bilinear [41]	36.9	-	-	-
FD-GAN [37]	40.8	-	-	-
DSR [7]	40.8	58.2	65.2	30.4
SFR [8]	42.3	60.3	67.3	32.0
PGFA [13]	51.4	68.6	74.9	37.3
Ours	**56.3**	**72.4**	**78.0**	**43.5**

TABLE III
Inference speed on Occluded-DukeMTMC. The inference time is the seconds per query.

Method	Time	Method	Time
PGFA$_{\text{w/o pose}}$ [13]	0.13s	PCB [2]	0.09s
PGFA$_{\text{w/ pose}}$ [13]	0.78s	DSR [7]	4.54s
Ours$_{\text{w/o pose}}$	0.14s	SFR [8]	4.76s
Ours$_{\text{w/ pose}}$	0.79s	-	-

Evaluation Metrics. For performance evaluation, we utilize the standard metrics as in most person re-id approaches, including the cumulative matching cure (CMC) and the mean Average Precision (mAP).

B. Implementation Details

We utilize AlphaPose [65], [66] pre-trained on the COCO dataset [54] as our human pose estimator. The threshold γ is set to 0.2. The backbone of our method is ResNet50 [53] and initialized by an ImageNet [67] pre-trained model. During the training procedure, the input image is resized to 384×128, following previous re-id approaches [2], [10]. We use random flipping to augment the training images. Besides, we randomly erase part of the input image to simulate the occlusion images [58], and the landmarks in the erased part are labeled as the occluded landmarks. Thus we can generate corresponding visible landmark vectors and enlarge the varieties. We use a batch size of 32, and train the model for 60 epochs. On the three large-scale datasets, *i.e.*, Occluded-DukeMTMC, Market-1501, and DukeMTMC-reID, the initial learning rate is set to 0.1 and reduced by a factor of 10 in the last 20 epochs. The parameter λ is 0.5. On the two small partial re-id datasets, Partial-REID and Partial-iLIDS, the learning rate is 0.02, reduced by a factor of 10. The initial λ is 0.9.

C. Comparison to the State-of-the-Art methods

Evaluations on Occluded-DukeMTMC. As shown in Table. II, the first group shows the approaches for the non-occluded re-id. The second group shows the methods of

TABLE IV
RESULTS ON PARTIAL-REID AND PARTIAL-ILIDS.

Method	Partial-REID		Partial_iLIDS	
	Rank-1	Rank-3	Rank-1	Rank-3
MTRC [62]	23.7	27.3	17.7	26.1
AMC+SWM [5]	37.3	46.0	21.0	32.8
DSR [7]	50.7	70.0	58.8	67.2
SFR [8]	56.9	78.5	63.9	74.8
VPM [6]	67.7	81.9	67.2	76.5
PGFA [13]	68.0	80.0	69.1	80.9
Ours	**72.5**	**83.0**	**70.6**	**81.3**

TABLE V
RESULTS ON MARKET-1501 AND DUKEMTMC-REID.

Method	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
SVDNet [3]	82.3	62.1	76.7	56.8
BoW+kissme [55]	44.4	20.8	25.1	12.2
PAN [50]	82.8	63.4	71.7	51.5
PAR [34]	81.0	63.4	-	-
Pedestrian[63]	82.0	63.0	-	-
DSR [7]	83.5	64.2	-	-
MultiLoss [15]	83.9	64.4	-	-
TripletLoss [29]	84.9	69.1	-	-
Adver Occluded [61]	86.5	78.3	79.1	62.1
MLFN [64]	90.0	74.3	81.0	62.8
PCB [2]	92.4	77.3	81.9	65.3
PGFA [13]	91.2	76.8	82.6	65.5
Ours	**92.7**	**81.3**	**86.2**	**72.6**

TABLE VI
THE EFFECTIVENESS OF THE POSE EMBEDDING, THE POSE-MASKED FEATURE BRANCH AND THE MATCHING STRATEGY.

Method	Rank-1	Rank-5	Rank-10	mAP
Ours$_{\text{w/o sim w/o pe}}$	51.4	68.6	74.9	37.3
Ours$_{\text{w/o pe}}$	53.2	69.4	75.3	40.5
Ours$_{\text{w/o sim}}$	54.1	69.6	75.2	40.3
Ours$_{\text{w/o pm}}$	55.0	70.5	76.7	41.5
Ours$_{\text{w/o matching}}$	51.2	62.4	73.4	41.2
Ours	**56.3**	**72.4**	**78.0**	**43.5**

using human pose information. The third group shows the partial re-id methods. PGFA [13] in the fourth group denotes our conference version for the occluded re-id. Our method surpasses all the previous methods by a large margin.

The partial feature branch of our method is similar to PCB [2]. The difference between PCB and our methods is that our method utilizes visible landmarks to encode the occlusion information into the feature representation, and use part features in the commonly visible parts for comparison during matching. The feature comparison in the visible shared region explicitly depresses the occlusion part in the occlusion images. Compared with PCB [2], our method surpasses it by +13.7% Rank-1 accuracy and +9.8% mAP, demonstrating the effectiveness of our proposed strategies.

Compared with our conference version, PGFA [13], this paper proposes the Pose-Embedded Feature Branch, which uses the visible landmark vector to generate the pose embedding. To generate larger varieties of the visible landmark vectors, we simulate the occlusion images in the training set. The result shows that adding the pose embedding improves Rank-1 accuracy from 51.4% to 56.3% (+4.9%), while improves mAP from 37.3% to 43.5% (+6.2%), which demonstrates the effectiveness of the Pose-Embedded Feature Branch.

We compare the inference speed of our method with the baseline method PCB [2], the partial re-id methods (DSR [7], and SFR [8]) and the conference version PGFA [13], as shown in Table. III. "w/ pose" or "w/o pose" indicate the method with or without the pose extracting. Table. III shows that our method is slightly slower than the conference version PGFA because of adding the Pose-Embedded Branch. Extracting the pose landmarks is time-consuming; thus, it is better to extract the landmarks in advance in practice. Our method is much faster than the partial re-id methods (DSR and SFR) because there is no time-consuming feature map matching during inference in our method.

Evaluations on Partial-REID and Partial-iLIDS. Some partial re-id approaches [62], [5], [7], [8], [6] are proposed for the partial re-id and evaluated on Partial-REID and Partial-iLIDS. To illustrate the effectiveness of our method on the partial re-id, we compare the results with these methods. We use the train set of Market-1501 for training, following previous methods [7], [8], [6]. As shown in Table. IV, our method surpasses previous partial re-id methods [62], [5], [7], [8], [6] on both Partial-REID and Partial-iLIDS datasets. Compared with our previous PGFA [13], adding the pose embedding improves the performance in these two datasets.

Evaluations on Market-1501 and DukeMTMC-reID. Table.V shows the results of our model on non-occluded re-id datasets, Market-1501 and DukeMTMC-reID. Our method achieves better results than the state-of-the-arts. The results show that our method can handle not only the occluded re-id but also the non-occluded re-id problem.

D. Ablation Studies

The Effectiveness of the Pose-Embedded Feature Branch. Compared with our conference version, PGFA [13], this paper proposes to use the visible landmark vector to generate the pose embedding. The pose embedding contains an informative knowledge of occlusions and is used to construct the pose-guided feature. To generate larger varieties of the visible landmark vectors, we simulate the occlusion images randomly. In Table. VI, Ours$_{\text{w/o sim w/o pe}}$ denotes our method without both data simulation and the pose embedding, which is the same as PGFA [13]. Ours$_{\text{w/o pe}}$ and Ours$_{\text{w/o sim}}$ denote our method without the pose embedding and data simulation, respectively. Results show that even without the data simulation, adding the pose embedding can improve mAP by +3.0%. Utilizing the data simulation can further improve mAP by +3.2%. Thus, both data simulation and the pose embedding take effects in our method.

The Effectiveness of the Pose-Masked Feature Branch. The Pose-Masked Feature Branch filter out the noise of occlusions by pose masks. In Table. VI, Ours$_{\text{w/o pm}}$ is our method without the pose-masked feature. Using the pose-masked feature improves Rank-1 from 55.0% to 56.3% and mAP from 41.5% to 43.5%, demonstrating its effectiveness.

The Effectiveness of the Commonly Visible Parts Matching Strategy. The commonly visible parts matching strategy

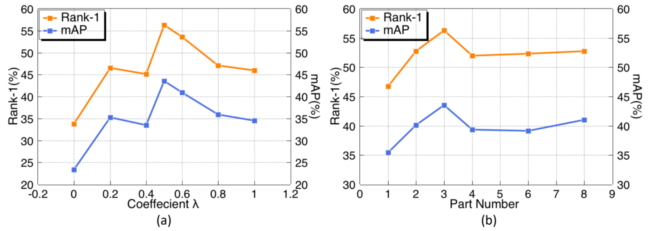

Fig. 5. (a) Varying Trade-off Coefficients λ. $\lambda = 0$ means only using the pose-guided feature. $\lambda = 1$ means only using the partial features with our matching strategy. (b) The impact of the part number p.

TABLE VII
COMPARISON ABOUT MULTIPLE GRANULARITIES.

Method	Rank-1	Rank-5	Rank-10	mAP
$p = 2, 3$	54.3	70.8	77.3	42.3
$p = 3, 4$	56.3	72.2	78.1	43.6
Ours ($p = 3$)	56.3	72.4	78.0	43.5

employs visible landmarks to select visible partial features. The distractive features from occlusion regions are filtered out. To illustrate the effectiveness of the matching strategy in testing, we evaluate our model using all the partial features for comparison. In Table. VI, Ours$_{\text{w/o matching}}$ is our method without the commonly visible parts matching strategy. The result without the matching strategy achieves 51.2% Rank-1 accuracy and 41.2% mAP. Thus, utilizing this strategy during matching improves Rank-1 by +5.0% and mAP by +2.3%.

Varying the Loss Coefficient. The final loss consists of two parts, $\mathcal{L}_{part}$ and $\mathcal{L}_{pose}$, corresponding to the impacts of partial features and the pose-guided feature. The coefficient λ balance the contributions of $\mathcal{L}_{part}$ and $\mathcal{L}_{pose}$. We conduct an ablation study on the coefficient λ, and λ grows from 0 to 1. $\lambda = 0$ means only the pose-guided branches take effect while $\lambda = 1$ means only the Partial Feature Branch takes effect. As shown in Fig. 5 (a), when $0 < \lambda < 1$, the retrieval performance is higher than $\lambda = 0$ or $\lambda = 1$. Thus, using a multiple-branch architecture performs better than only one of these branches. When $\lambda = 0.5$, our method achieves the best performance.

The Impact of the Part Number p. p determines the granularity of the part features. $p = 1$ means that our method uses a global feature instead of partial features. In Fig. 5 (b), the retrieval performance is higher when $p > 1$ than $p = 1$, illustrating that the partition strategy is helpful to generate more robust features and split the occlusions and the target person. When $p = 3$, our method achieves the best performance. When p is too large ($p > 3$), some partial feature is too small and doesn't contain the visible landmarks, although it is the visible part. Thus, the informative partial feature of the target person is filtered out. We also compare our method with methods [31], [32] that horizontally partition the feature map with multiple granularities. As shown in Table. VII, $p = 2, 3$ means the feature map is partitioned into two parts and three parts simultaneously, with a total of five classifiers to optimize. $p = 3, 4$ means the feature map is partitioned into three parts and four parts. Results show that multiple granularities achieve similar performance in our setting. One possible reason is that some partial features of multiple granularities are filtered out by landmarks when facing the occlusion problem and do not take effect.

The Impact of the gaussian heatmap. Fig. 6 (a) shows the impact of the parameter σ of the gaussian heatmap. When σ is small, the visible area in the generated Gaussian heatmaps is small (Only the small area around the visible landmarks is non-zero). Thus the useful information from the Pose-Masked Branch is limited, and the performance is affected. When σ is large, the heatmaps cannot filter out the occluded parts, since most part of the heatmap is close to 1. We choose $\sigma = 20$ in this paper for the best performance.

TABLE VIII
COMPARISON OF HUMAN POSE ESTIMATORS.

Method		Rank-1	Rank-5	mAP
AlphaPose [65]	Ours$_{\text{w/o sim w/o pose}}$	51.4	68.6	37.3
	Ours$_{\text{w/o sim}}$	54.1	69.6	40.3
	Ours	56.3	72.4	43.5
OpenPose [68]	Ours$_{\text{w/o sim w/o pose}}$	49.1	66.7	35.3
	Ours$_{\text{w/o sim}}$	52.3	67.6	38.5
	Ours	54.3	71.2	41.2

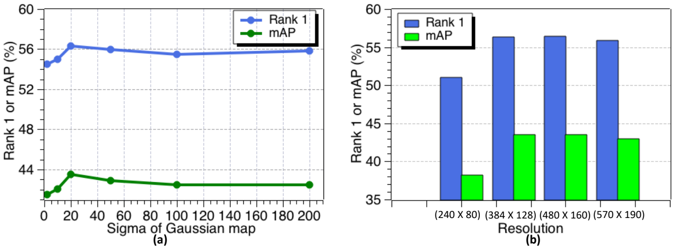

Fig. 6. (a) Impact of the sigma of the gaussian heatmaps. (b) Impact of the input resolution.

Fig. 6 (b) shows the impact of the heatmap resolution. When enlarging the input resolution from 384×128 to 480×160 and 570×190, the models achieve similar performance. This indicates that enlarging the heatmaps with the same σ does not affect the re-id performance. When the resolution is reduced, the performance drops because of the information drop.

The Impact of the Pose Estimation Algorithm. Previous experiments use AlphaPose [65] as the human pose estimator. To evaluate the sensitiveness of our method to the human pose estimator, we evaluate our method using OpenPose [68]. As shown in Table. VIII, using two pose estimators achieve similar performance, illustrating that our method is not sensitive to the pose estimation algorithms on Occluded-DukeMTMC.

E. Visualization

Fig. 7 shows the visualization of the pose masks generated by the visible landmarks. The pose masks focus on the visible parts of the target person in the image, while the occlusions are depressed. Fig. 8 shows some evaluation results of PCB [2] and our method on Occluded-DukeMTMC. The PCB method tends to introduce the distractive information of the occlusions, and a probe occluded by an obstacle is easy to retrieve incorrect images with a similar obstacle. Compared with PCB, our method filters out the information of the occlusions and achieves better performance.

VI. CONCLUSION

This paper focuses on the occluded person re-id problem. In our setting, all images in the query set are occluded, while the gallery set contains both occluded and non-occluded images,

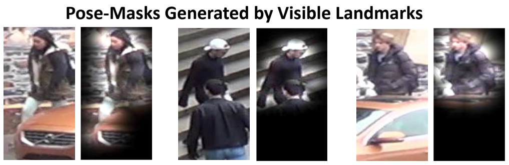

Fig. 7. The visualization of the pose masks generated by visible landmarks.

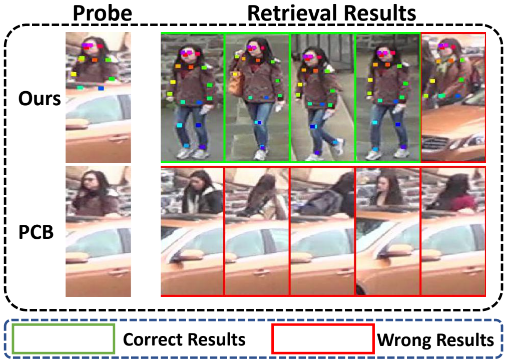

Fig. 8. Visulization of the retrieval results on PCB [2] and our method.

which is more challenging and practical. We propose to benefit the occluded re-id by pose landmarks in three aspects. First, we use the spatial location information of the visible landmarks to filter the noise of occlusion regions. Second, we use visible landmarks to generate the pose embedding, which is used as the channel gates to re-calibrate the channel features. Third, in testing, the commonly visible part features are used for comparison. Besides, we construct a large-scale occluded re-id dataset, Occluded-DukeMTMC.

References

[1] X. Yang, P. Zhou, and M. Wang, "Person reidentification via structural deep metric learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 2987–2998, 2018.

[2] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[3] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3800–3808.

[4] L. An, X. Chen, S. Yang, and X. Li, "Person re-identification by multi-hypergraph fusion," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2763–2774, 2016.

[5] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4678–4686.

[6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 393–402.

[7] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7073–7082.

[8] L. He, Z. Sun, Y. Zhu, and Y. Wang, "Recognizing partial biometric patterns," *arXiv preprint arXiv:1810.07399*, 2018.

[9] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[10] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8450–8459.

[11] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 542–551.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724*, 2017.

[16] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5177–5186.

[17] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE transactions on neural networks and learning systems*, 2019.

[18] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[19] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 9387–9399, 2020.

[20] W. Zhang, X. He, W. Lu, H. Qiao, and Y. Li, "Feature aggregation with reinforcement learning for video-based person re-identification," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 12, pp. 3847–3852, 2019.

[21] L. Wu, Y. Wang, L. Shao, and M. Wang, "3-d personvlad: Learning deep global representations for video-based person reidentification," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3347–3359, 2019.

[22] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 728–739, 2020.

[23] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2872–2881, June 2019.

[24] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020.

[25] H. Fan and Y. Yang, "Person tube retrieval via language description," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 754–10 761.

[26] L. Wu, Y. Wang, J. Gao, M. Wang, Z. J. Zha, and D. Tao, "Deep coattention-based comparator for relative representation learning in person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2020.

[27] Z. zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person re-identification," in *TOMM*, 2017.

[28] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, "Unsupervised person re-identification via cross-camera similarity exploration," *IEEE Transactions on Image Processing*, vol. 29, pp. 5481–5490, 2020.

[29] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[30] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[31] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8295–8302.

[32] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[33] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[34] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[35] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[36] H. Wang, L. Jiao, S. Yang, L. Li, and Z. Wang, "Simple and effective: Spatial rescaling for person reidentification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2020.

[37] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in neural information processing systems*, 2018, pp. 1222–1233.

[38] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[39] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[40] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[41] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[42] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[43] L. Yang, Q. Song, Y. Wu, and M. Hu, "Attention inspiring receptive-fields network for learning invariant representations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1744–1755, 2019.

[44] Z. Liang and J. Shen, "Local semantic siamese networks for fast tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 3351–3364, 2019.

[45] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 459–474.

[46] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3516–3527, 2019.

[47] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, "Occlusion-aware real-time object tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 763–771, 2016.

[48] X. Dong, J. Shen, W. Wang, L. Shao, H. Ling, and F. Porikli, "Dynamical hyperparameter optimization via deep reinforcement learning in tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[49] J. Shen, D. Yu, L. Deng, and X. Dong, "Fast online tracking with detection refinement," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 162–173, 2017.

[50] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[51] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[52] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[54] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[55] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[56] Q. Yu, X. Chang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching," *arXiv preprint arXiv:1711.08106*, 2017.

[57] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[58] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[59] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[60] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[61] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[62] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 35, no. 5, pp. 1193–1205, 2012.

[63] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2018.

[64] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2109–2118.

[65] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[66] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose Flow: Efficient online pose tracking," in *BMVC*, 2018.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[68] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

Jiaxu Miao received the B.E. degree from Shanghai Jiao Tong University, China, in 2015. He is currently a Ph.D. candidate in the Australian Artificial Intelligence Institute, University of Technology Sydney, Australia. His research interests are person re-identification and video segmentation.

Yu Wu is currently a Ph.D. candidate in the Australian Artificial Intelligence Institute, University of Technology Sydney, Australia. He received the B.E. degree from Shanghai Jiao Tong University, China, in 2015. His research interests are video analysis and multi-modal perception.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with the University of Technology Sydney, Australia. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video content understanding.