Semantic-Aware Occlusion-Robust Network for Occluded Person Re-Identification

Xiaokang Zhang, Yan Yan, *Member, IEEE,* Jing-Hao Xue, Yang Hua, and Hanzi Wang, *Senior Member, IEEE*

***Abstract*—In recent years, deep learning-based person re-identification (Re-ID) methods have made significant progress. However, the performance of these methods substantially decreases when dealing with occlusion, which is ubiquitous in realistic scenarios. In this paper, we propose a novel semantic-aware occlusion-robust network (SORN) that effectively exploits the intrinsic relationship between the tasks of person Re-ID and semantic segmentation for occluded person Re-ID. Specifically, the SORN is composed of three branches, including a local branch, a global branch, and a semantic branch. In particular, the local branch extracts part-based local features, and the global branch leverages a novel spatial-patch contrastive loss (SPC) to extract occlusion-robust global features. Meanwhile, the semantic branch generates a foreground-background mask for a pedestrian image, which indicates the non-occluded areas of the human body. The three branches are jointly trained in a unified multi-task learning network. Finally, pedestrian matching is performed based on the local features extracted from the non-occluded areas and the global features extracted from the whole pedestrian image. Extensive experimental results on a large-scale occluded person Re-ID dataset (i.e., Occluded-DukeMTMC) and two partial person Re-ID datasets (i.e., Partial-REID and Partial-iLIDS) show the superiority of the proposed method compared with several state-of-the-art methods for occluded and partial person Re-ID. We also demonstrate the effectiveness of the proposed method on two general person Re-ID datasets (i.e., Market-1501 and DukeMTMC-reID).**

***Index Terms*—Person re-identification, occlusion, semantic segmentation, multi-task learning.**

I. Introduction

PERSON re-identification (Re-ID) aims to retrieve a query pedestrian image from a gallery collected across several non-overlapping cameras. It has attracted considerable attention because of its variety of applications, such as person search in video surveillance and person tracking. During the last few decades, significant progress has been made in person Re-ID. However, it is still a challenging task confronted with many challenges, such as various background clutter and complex pedestrian appearance variations caused by different poses, camera views, and illuminations.

Over the past few years, deep learning-based methods have dominated the person Re-ID research due to their outstanding performance on retrieval accuracy. Some methods [1], [2] employ deeply learned global features, while a large number of methods [3]–[5] resort to taking advantage of predefined rigid regions to extract local features. A few methods [6]–[11] employ external auxiliary models (such as pose estimation models) for feature extraction and alignment.

X. Zhang, Y. Yan, and H. Wang are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: xiaokangz@stu.xmu.edu.cn; yanyan@xmu.edu.cn; hanzi.wang@xmu.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (e-mail: jinghao.xue@ucl.ac.uk).

Y. Hua is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK (e-mail: y.hua@qub.ac.uk).

Fig. 1. Examples of occluded pedestrian images from the Occluded-DukeMTMC [12], Partial-iLIDS [13], and Partial-REID [14] datasets.

In realistic person Re-ID scenarios, occlusion is ubiquitous, especially in crowded public places. For example, pedestrians are often occluded by various obstacles, such as cars, umbrellas, pillars, and other pedestrians. Several examples are shown in Fig. 1. Occlusion may introduce severe disturbance to the trained models, thus resulting in difficulty in learning robust feature representations by conventional deep learning-based methods. As a result, the performance of these methods substantially decreases when dealing with the occlusion problem (see Fig. 2 for an illustrative example).

Recently, several methods [14]–[18] have been proposed to address the problem of partial person Re-ID, where the query images are occluded by various obstacles while the gallery images are non-occluded. However, these methods usually manually crop the occluded areas of the query images and then use the non-occluded areas for retrieval to alleviate the disturbance caused by occlusion. Clearly, such a manner is not practical since both the query and the gallery may contain occluded images in real-world scenarios. Moreover, the manual cropping process is not efficient.

Different from the problem of partial person Re-ID, the problem of occluded person Re-ID is more challenging and practical, where both the query and the gallery can contain occluded images. In addition, the manual cropping operation is not allowed, to avoid human bias. Representative works for handling occluded person Re-ID are the pose-guided feature alignment (PGFA) method [12] and the high-order person Re-ID (HOReID) method [36]. PGFA relies on human landmarks

to disentangle informative information in the non-occluded areas from occluded images. HOReID learns high-order relation information for extracting discriminative features and topology information for performing robust alignment based on a pose estimator. However, PGFA and HOReID train the tasks of person Re-ID and pose estimation separately, and they ignore the close connection between the two tasks. Moreover, the performance of these methods depends heavily on the accuracy of the pose estimation model. As a result, these methods might not be able to extract effective features based on these landmarks when the generated human landmarks are not accurate due to occlusion, thus leading to a performance decrease.

In this paper, to solve the challenging problem of occluded person Re-ID, we propose a deep model called the semantic-aware occlusion-robust network (SORN) by taking advantage of semantic segmentation and exploring the important relationship between the tasks of occluded person Re-ID and semantic segmentation. Specifically, the SORN contains a local branch, a global branch, and a semantic branch. Inspired by the part-based convolutional baseline (PCB) method [3], the local branch evenly partitions the feature map into several horizontal stripes and adopts average pooling to obtain effective local features. Meanwhile, the global branch extracts occlusion-robust global features, where a novel spatial-patch contrastive (SPC) loss is proposed to enforce the global features to encode discriminative local information. The semantic branch plays the role of semantic segmentation, which can indicate the non-occluded areas of the human body. Since current person Re-ID datasets do not have the semantic labels for pedestrian images, we adopt a pretrained semantic segmentation model and label smoothing to predict the semantic labels. In this way, the semantic branch can be jointly trained with the local and global branches in an end-to-end manner. Based on the trained model, pedestrian matching is performed by using the local features extracted from the non-occluded areas and the global features extracted from the whole pedestrian image.

Fig. 2 gives the top 5 retrieval results obtained by the PCB method and the proposed method for a query image in the Occluded-DukeMTMC dataset. By considering the tasks of occluded person Re-ID and semantic segmentation in a unified multi-task learning network, the proposed method significantly alleviates the disturbance caused by occlusion and thus correctly retrieves the person of interest.

In summary, the major contributions of our work are summarized as follows:

- We propose a novel semantic-aware occlusion-robust network (SORN), which consists of a local branch, a global branch, and a semantic branch, for occluded person Re-ID. By incorporating semantic segmentation into occluded person Re-ID, the intrinsic relationship between these two tasks is fully exploited. Moreover, the negative effect of occluded areas and background clutter is effectively alleviated.
- We propose a novel spatial-patch contrastive (SPC) loss to extract occlusion-robust global features in the global branch. The global features are highly discriminative and robust to occlusion. More importantly, the global features can be reliably used for occluded pedestrian matching, especially when the occluded areas are different for the query and gallery images.
- Experimental results demonstrate that the proposed method performs favorably against state-of-the-art methods on the problem of occluded person Re-ID. Moreover, we also show the effectiveness of the proposed method on the problems of partial person Re-ID and general person Re-ID.

Fig. 2. The top 5 retrieval results obtained by PCB and our proposed SORN method for a query image in the Occluded-DukeMTMC dataset. The images with green and red borders denote the correct and incorrect retrieval results, respectively. The patch in the query image with a blue border indicates that the region is not severely occluded, while that with a yellow border represents that the region is occluded by an obstacle.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III first introduces the overall framework of the proposed method and then illustrates the three branches and the pedestrian matching strategy. Section IV presents the experimental results. Finally, Section V draws the conclusion.

II. Related Work

In this section, we briefly review the related work on general person Re-ID methods, partial person Re-ID methods, and occluded person Re-ID methods.

A. General Person Re-ID Methods

Recently, with the development of deep learning, the performance of general person Re-ID methods has received significant improvements [19]. Existing general person Re-ID methods mainly focus on two aspects: (1) learning discriminative feature representations [3], [7], [8], [20]–[25]; and (2) learning effective metrics [1], [26]–[30].

On the one hand, for learning discriminative feature representations, many methods attempt to extract fine-grained local features. For instance, Sun *et al.* [3] propose the PCB method, which extracts local features by dividing the feature map into uniform patches. PCB also introduces refined part pooling (RPP) to further boost the performance of person Re-ID. Lei *et al.* [7] propose to perform person Re-ID based on the semantic region representation and a mapping space topology constraint,

where the semantic information is employed to alleviate the misalignment problem (caused by viewpoint changes and pose variations). Zheng *et al.* [20] develop a pedestrian alignment network (PAN) to jointly align pedestrian images and learn discriminative features. Shen *et al.* [21] propose sharp attention networks (SAN) to generate sharp attention masks, which can assertively select subtle visual structures by sampling from convolutional features. The above methods explicitly address various challenges (such as the misalignment problem and attention mask generation) for general person Re-ID.

On the other hand, for learning effective metrics, most methods adopt some regularization terms or constraints to enforce the person Re-ID model to learn discriminative feature embedding. Representative methods include triplet loss [27], hard-aware point-to-set (HAP2S) loss [28], quadruplet loss [1], group similarity loss [29], and hinge loss [30], where the design of loss functions and the strategy of hard sample mining play critical roles. Note that these methods are developed to address the problem of general person Re-ID. Therefore, when the pedestrian is occluded by obstacles in the image, the feature representations extracted by these methods tend to be noisy, thus leading to a significant performance decrease.

Recently, some pose-guided methods use human landmarks to locate the human body and alleviate the misalignment problem for matching pedestrian images. For example, Su *et al.* [6] propose a pose-driven deep convolutional (PDC) model to learn effective feature representations and adaptive similarity measures. PDC explicitly leverages human landmarks to handle the problem of person Re-ID under pose variations. Ge *et al.* [9] propose a pose-guided feature distilling generative adversarial network (FD-GAN) model, which makes use of human landmarks to generate pedestrian images. A novel same-pose loss is integrated into the FD-GAN model to simultaneously learn identity-related and pose-unrelated representations. Zheng *et al.* [11] propose to address the pedestrian misalignment problem by introducing pose invariant embedding (PIE) as a pedestrian descriptor. Suh *et al.* [31] propose a two-stream network that can generate appearance and body-part feature maps.

Some mask-guided methods employ person masks to extract features for person Re-ID. For instance, Kalayh *et al.* [8] propose the SPReID method to extract body-part features by using an additionally trained human parsing model. Song *et al.* [32] use the source image and the corresponding binary segmentation mask as inputs to extract discriminative features that are invariant to background clutter. Qi *et al.* [24] adopt both the source image and the masked image as the network inputs, where a multi-layer fusion scheme and a ranking loss are developed to fuse the different levels of features and optimize the network, respectively. The mask-guided methods can extract aligned local features and focus on foreground areas by exploiting the results from semantic segmentation. However, these methods cannot effectively perform occluded person Re-ID, where the occluded areas can be different for the query and the gallery.

The pose-guided or mask-guided methods resort to external models (e.g., pose estimation or human parsing models) for general person Re-ID. Although these methods can alleviate occlusion to some extent, they depend heavily on accurate pose estimation (or human parsing). Moreover, the pose estimation (or human parsing) model and the person Re-ID model are usually independently trained since the ground-truth human landmarks (or semantic labels) are not available in the person Re-ID datasets. As a result, these methods ignore the inherent dependency between the tasks of pose estimation (or human parsing) and person Re-ID. In contrast, our method jointly trains the tasks of occluded person Re-ID and semantic segmentation in a unified multi-task learning network. In particular, we predict the semantic labels of the training dataset by using the DANet model [33] and adopt label smoothing to prevent overfitting (caused by some false semantic labels predicted by DANet) on the training dataset. In this way, our method effectively explores the intrinsic relationship between the two tasks. More importantly, semantic segmentation is helpful to reduce the negative influence of occluded areas and background clutter, thus leading to performance improvements.

B. Partial Person Re-ID Methods

Partial person Re-ID aims to match the query partial pedestrian image against the full-body pedestrian images in the gallery. Zheng *et al.* [14] propose a local patch-level matching method to explicitly model the ambiguity of the occlusion patterns. They also introduce a global part-based matching model to encode the spatial layout information. He *et al.* [15] propose to leverage a fully convolutional network (FCN) to generate fixed-sized spatial feature maps and use deep spatial feature reconstruction (DSR) to match pedestrian images.

Recently, He *et al.* [16] develop a dictionary learning-based spatial feature reconstruction (SFR) method to match different sized feature maps for partial person Re-ID. Luo *et al.* [17] propose a spatial transformer network (STN) to sample a semantic patch from the holistic image to match the partial image and jointly train the STN module and the person ReID module in a two-step procedure. Sun *et al.* [18] propose a visibility-aware part model (VPM) that automatically identifies the visibility of regions in a partial image based on self-supervision.

The above methods usually manually crop the occluded areas of the query pedestrian image. However, the manual cropping process is not practical and efficient. Different from these methods, the proposed method addresses a more general scenario (both the query images and the gallery images can contain occlusion) and does not require manual cropping.

C. Occluded Person Re-ID Methods

Although occlusion is a major challenge in person Re-ID, there are only a few studies on the problem of occluded person Re-ID. Zhuo *et al.* [34] propose an attention framework of person body (AFPB) method for occluded person Re-ID. The method employs an occlusion simulator (OS) to generate artificial occluded images by randomly adding background patches to full-body pedestrian images. However, such a method cannot effectively deal with the case in which both the query and the gallery contain occlusion. He *et al.* [35]

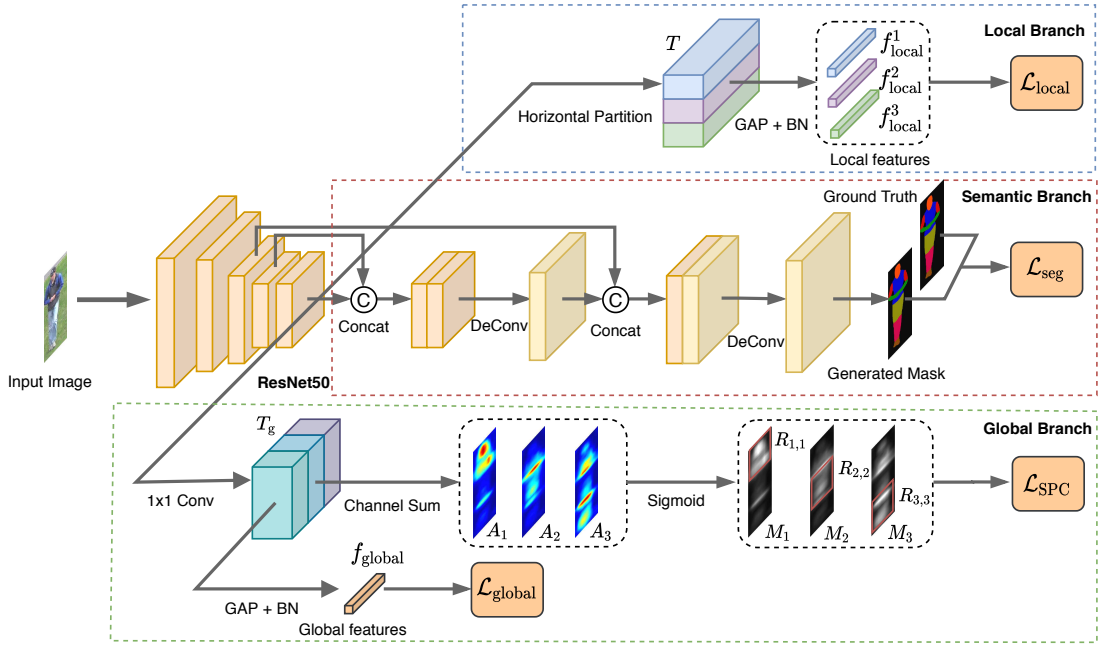


Fig. 3. The overall framework of our proposed SORN for occluded person Re-ID. Our method consists of a ResNet50 backbone, a local branch (Section III-B), a global branch (Section III-C), and a semantic branch (Section III-D). The three branches are effectively combined and jointly trained in an end-to-end way. Here, '©' represents the concatenation operation. The number of patches P and the number of channel sets N are both set to 3 for simplicity.

employ an FCN and pyramid pooling to extract spatial pyramid features, and use foreground-aware pyramid reconstruction (FPR) to calculate the matching scores between the query image and the gallery. However, to effectively reconstruct the query image, the images in the gallery are required to be non-occluded. Wang *et al.* [36] leverage a convolutional neural network (CNN) backbone and a pose estimation model to extract local features. An adaptive direction graph convolutional (ADGC) layer and a cross-graph embedded-alignment (CGEA) layer are developed to embed the relation information and the topology information, respectively. Zhuo *et al.* [37] design a teacher-student learning framework to learn an occlusion-robust model. Miao *et al.* [12] propose the PGFA method for occluded person Re-ID. PGFA leverages an additionally trained pose estimator to generate pose landmarks, which indicate the occluded areas.

In this paper, we integrate semantic segmentation with occluded person Re-ID in a unified multi-task learning network, where a semantic branch, a global branch, and a local branch are jointly trained in an end-to-end manner. Therefore, our method fully exploits the semantic information to enable the model to pay attention to the non-occluded areas. Moreover, compared with PGFA [12] that depends on accurate pose estimation, our method is more tolerant to small segmentation errors by taking the semantic information into account.

III. Proposed Method

In this section, we present the details of our proposed SORN method for occluded person Re-ID. An overview of the proposed method is introduced in Section III-A. Each component of the proposed method is described in detail from Section III-B to Section III-E. Finally, the discussions between our proposed method and several state-of-the-art methods are given in Section III-F.

A. Overview

The proposed SORN method contains four main components, including a feature extraction backbone, a local branch, a global branch, and a semantic branch. The overall framework of the proposed method is shown in Fig. 3.

The pedestrian images are first input to a ResNet50 backbone [38] to generate a 3D feature map $\mathbf{T} \in \mathbb{R}^{h \times w \times c}$, where h, w, and c denote the height, width, and number of channels, respectively. Similar to [3], the spatial down-sampling operation in the last layer of ResNet50 is removed to obtain the feature map with a higher spatial resolution.

Next, the feature map $\mathbf{T}$ is fed into a local branch, a global branch, and a semantic branch. For the local branch, $\mathbf{T}$ is divided into several patches in the horizontal direction, and local features are then obtained by applying global average pooling. For the global branch, occlusion-robust global features, which can preserve local details of the pedestrian image, are extracted based on the proposed spatial-patch contrastive (SPC) loss. For the semantic branch, the feature maps from different levels of the ResNet50 backbone are used as the inputs, and several deconvolutional layers are employed to generate a foreground-background mask. Finally, the three branches are jointly trained in an end-to-end manner.

For the testing stage, the local features are selected from the non-occluded areas based on the generated foreground-background mask. Then, pedestrian matching is performed according to the global features and the selected local features.

B. Local Branch

For the local branch, we first evenly partition the feature map $\mathbf{T}$ into P patches in the horizontal direction. Note that each patch has a fixed position based on prior knowledge about the human body structure. Then, the local features are obtained by employing a global average pooling (GAP) layer followed by a batch normalization (BN) layer for each patch. In this way, the local branch can extract fine-grained local features. We denote these local features as $\{\mathbf{f}_{local}^p\}_{p=1}^P$, where $\mathbf{f}_{local}^p$ represents the local features extracted from the p-th patch and P is the number of patches. Finally, each local feature representation is fed into a fully-connected layer and a softmax layer to predict the probability $\mathbf{y}^p$ of a person ID. Here, $\mathbf{y}^p = [y_1^p, \cdots, y_J^p]$, and J is the number of person IDs.

The classification loss of the local branch is formulated according to the sum of the cross-entropy loss, that is,

$$\mathcal{L}_{local} = -\sum_{p=1}^{P}\sum_{j=1}^{J} q_j \log(y_j^p), \tag{1}$$

where q_j denotes the label indicator. That is, $q_j = 1$ if j is equal to the ground-truth label, and $q_j = 0$, otherwise.

Note that PCB [3] and VPM [18] also employ the local branch. However, different from these methods, we add a BN layer after the GAP layer. Such a manner guarantees that the gradients are more predictive and stable, which can avoid overfitting during training [39].

C. Global Branch

For the global branch, we first apply a 1×1 convolutional layer to the feature map $\mathbf{T}$ so that the number of channels is extended from c to Nc' (where N is the number of channel sets and c' is the reduced number of channels in each set). Hence, a 3D feature map $\mathbf{T}_g \in \mathbb{R}^{h \times w \times Nc'}$ is obtained. Then, the Nc' channels of $\mathbf{T}_g$ are uniformly divided into N different sets of channels. Obviously, the n-th set of channels refers to the channels from the $[(n-1)c'+1]$-th channel to the nc'-th channel, where $n \in \{1, \cdots, N\}$.

The 2D feature map of a single channel only contains weak semantic information and can easily be affected by disturbance [40]. Therefore, we combine the 2D feature maps of the n-th set of channels and obtain a 2D aggregation map $\mathbf{A}_n \in \mathbb{R}^{h \times w}$ as follows:

$$\mathbf{A}_n = \sum_{i=(n-1)c'+1}^{nc'} \mathbf{T}_g^i, \tag{2}$$

where $\mathbf{T}_g^i$ denotes the i-th 2D feature map within $\mathbf{T}_g$.

To ensure the stability of the training procedure, we adopt a normalization step to constrain the range of $\mathbf{A}_n$ to $[0, 1]$. Specifically, we obtain the average value $\overline{A_n}$ of the aggregation map $\mathbf{A}_n$, which is defined as

$$\overline{A_n} = \frac{\sum_{(x,y)\in\mathbf{A}_n} \mathbf{A}_n(x, y)}{h \times w}, \tag{3}$$

where (x, y) denotes the spatial location in $\mathbf{A}_n$.

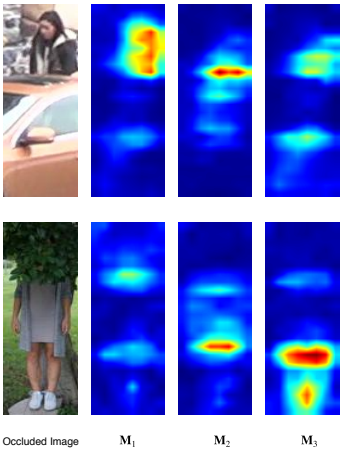


Fig. 4. Visualization of the normalized aggregation maps ($N = 3$) for two occluded images in the Occluded-DukeMTMC dataset. $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ denote the aggregation maps from the first, second, and third sets of channels, respectively.

Next, we employ a sigmoid function to normalize each element of $\mathbf{A}_n$ as

$$\mathbf{M}_n(x, y) = \text{sig}(\mathbf{A}_n(x, y) - \overline{A_n}), \tag{4}$$

where $\mathbf{M}_n$ denotes the normalized aggregation map. $\text{sig}(z) = 1/(1 + e^{-z})$ is the sigmoid function, which maps elements in $\mathbf{A}_n$ with values larger than $\overline{A_n}$ closer to 1, while those smaller than $\overline{A_n}$ closer to 0.

For general person Re-ID methods, the global features that capture the full receptive field can be used to discriminate different pedestrians. However, when dealing with occluded person Re-ID, the global features extracted from these methods usually contain noisy information due to the occluded areas. Therefore, for occluded person Re-ID, the desired global features are expected to exploit the information mainly from the non-occluded areas. Moreover, they should be able to effectively encode the context information to address the misalignment problem.

To achieve this goal, we uniformly divide each normalized aggregation map into N rectangular patches in the horizontal direction (since there are N normalized aggregation maps in total) and enforce each normalized aggregation map to activate a specific patch. Therefore, we propose a novel spatial-patch contrastive (SPC) loss as follows:

$$\mathcal{L}_{SPC} = \sum_{n=1}^{N}\Big(\sum_{(x,y)\in\mathbf{M}_n} \mathbf{M}_n(x, y) - \sum_{(x,y)\in\mathbf{R}_{n,n}} \mathbf{M}_n(x, y)\Big). \tag{5}$$

Here, $\mathbf{R}_{n,l}$ denotes the l-th rectangular patch of the n-th normalized aggregation map. Thus, $\mathbf{R}_{n,n}$ represents the n-th patch of the n-th normalized aggregation map.

In Eq. (5), we subtract the responses of $\mathbf{R}_{n,n}$ from those of $\mathbf{M}_n$. Therefore, the SPC loss minimizes the responses of $\mathbf{R}_{n,1} + \cdots + \mathbf{R}_{n,n-1} + \mathbf{R}_{n,n+1} + \cdots + \mathbf{R}_{n,N}$ for the n-th normalized aggregation map. In this way, by minimizing the SPC loss, different normalized aggregation maps focus on different patches (i.e., $\mathbf{M}_n$ focuses on the n-th patch $\mathbf{R}_{n,n}$).

Fig. 4 visualizes the normalized aggregation maps based on the SPC loss for two occluded pedestrian images in the Occluded-DukeMTMC dataset. We can see that each aggregation map focuses on a specific patch. In particular, when a patch is occluded (e.g., a woman is occluded by a car and a woman is occluded by a tree in the first and second rows of Fig. 4, respectively), the corresponding aggregation map (e.g., the 4th and 2nd images in the first and second rows of Fig. 4, respectively) suppresses the disturbance information in this patch. Therefore, the global branch can concentrate on the non-occluded areas of the pedestrian images, which makes the global features robust to occlusion and highly discriminative. Furthermore, the global features capture the full receptive field of the input image, which is beneficial for alleviating the misalignment problem.

Similar to the local branch, we also use global average pooling on $\mathbf{T}_g$ to extract the global features $\mathbf{f}_{global}$ and then obtain the ID prediction probability distribution $\mathbf{g} = [g_1, \cdots, g_J]$ by applying a fully-connected layer and a softmax layer. The classification loss of the global branch can be formulated as follows:

$$\mathcal{L}_{global} = -\sum_{j=1}^{J} q_j \log(g_j), \tag{6}$$

where q_j is the label indicator.

D. Semantic Branch

For occluded person Re-ID, the local features extracted from occluded areas contain the disturbance information. Therefore, it is of great importance to exclude disturbances in occluded areas. Existing methods [8] [12] rely on external models (such as the human parsing model or the pose estimation model) to indicate the occluded areas. However, these external models are usually independently trained, which ignores the relationship between the tasks of occluded person Re-ID and pose estimation (or human parsing).

In this paper, we use a semantic branch for detecting the occluded areas of pedestrian images in a unified network. The benefits of adopting the semantic branch are twofold. 1) By incorporating semantic segmentation into occluded person Re-ID, we expect that the model can focus on the non-occluded areas rather than background clutter or occluded areas. Such a manner enables the trained model to effectively improve the capability of distinguishing the non-occluded foreground from the background. As a result, the intrinsic relationship between occluded person Re-ID and semantic segmentation is fully exploited. Thus, the joint training of semantic branch and the other two branches (i.e., the global and local branches) helps to improve the performance of occluded person Re-ID. 2) For pedestrian matching, we capitalize on the generated foreground-background masks to determine the non-occluded areas and extract the local features from these areas. In this way, the disturbance in the occluded areas can be removed for more accurate retrieval (Section III-E).

The lower layers of the convolutional networks contain sufficient spatial information, while the upper layers of the networks encode high-level semantic information. Therefore, we leverage the feature maps from the different levels of the ResNet50 backbone as the inputs of the semantic branch to simultaneously exploit both spatial and semantic information. The architecture of our semantic branch is shown in Fig. 3.

Specifically, the output feature maps (with a size of $h \times w$) from the $3rd$ and $4th$ residual blocks of ResNet50 are concatenated as the input, which is fed into a 3×3 deconvolutional layer with stride 2, followed by batch normalization and ReLU. Then, the output feature map (with a size of $2h \times 2w$) from the $2nd$ residual block of ResNet50 and the output from the previous layer are concatenated and fed into another 3×3 deconvolutional layer with stride 2, followed by batch normalization and ReLU. Finally, a 1×1 convolutional layer is employed to classify each pixel of the final feature map (with a size of $4h \times 4w$) into K semantic classes. In this paper, eight semantic classes (including head, torso, upper arm, lower arm, upper leg, lower leg, foot, and background) are considered. The output of the convolutional layer is the prediction probability $\mathbf{s} = [s_1, \cdots, s_K]$ of semantic classes.

The loss of the semantic branch is defined as follows:

$$\mathcal{L}_{seg} = \sum_{k=1}^{K} -q_k \log(s_k). \tag{7}$$

Here, $q_k = \delta_{k,t}$ is the label distribution, where t is the ground-truth label. $\delta_{k,t}$ denotes the Dirac delta function, which is equal to 1 if $k = t$, and 0 otherwise.

It is worth noting that current person Re-ID datasets usually do not provide ground-truth semantic labels. In this paper, instead of relying on expensive manual labeling, we predict the semantic labels of the training dataset by using the DANet model [33] (i.e., a semantic segmentation model trained on the DensePose-COCO dataset [41]). As a result, some false semantic labels exist in the training dataset. These false semantic labels may lead to overfitting of the semantic branch since each training sample is assigned a full probability to the ground-truth label (predicted by DANet) according to Eq. (7). To overcome the above issue, we further adopt label smoothing (LS) proposed in [42] as the model regularization.

Generally, LS encourages the model to be less confident about the ground-truth label. Specifically, for a training sample with the ground-truth label t, the label distribution q_k is changed as

$$q_k = (1 - \epsilon)\delta_{k,t} + \epsilon u_k, \tag{8}$$

where ϵ is the label smoothing parameter. Eq. (8) can be viewed as a weighted combination of the original ground-truth label distribution $\delta_{k,t}$ and the fixed label distribution u_k. For simplicity, u_k uses the uniform distribution, i.e., $u_k = \frac{1}{K}$.

Thus, q_k can be reformulated as

$$q_k = \begin{cases} 1 - \frac{K-1}{K}\epsilon & k = t \\ \frac{\epsilon}{K} & k \neq t, \end{cases} \tag{9}$$

where ϵ is empirically set to 0.10 in all the experiments.

Finally, the overall loss of our SORN method can be formulated as

$$\mathcal{L} = \mathcal{L}_{local} + \lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{SPC} + \lambda_3 \mathcal{L}_{seg}, \tag{10}$$

where $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$ denote the classification losses of the local branch and the global branch, respectively. $\mathcal{L}_{SPC}$ denotes the proposed SPC loss. $\mathcal{L}_{seg}$ denotes the semantic segmentation loss. λ_1, λ_2 and λ_3 are the regularization parameters to balance different losses.

Based on Eq. (10), we can jointly train the three branches in an end-to-end manner.

E. Pedestrian Matching

For pedestrian matching, we consider the background label and the foreground label of the human body. That is, we aggregate all the semantic labels of the human body (seven classes in total) from the semantic branch as one foreground category. In this way, we can obtain a binary foreground-background mask $\mathbf{E} \in \mathbb{R}^{h' \times w'}$, where $h' = 4h$ and $w' = 4w$ respectively denote the height and width of the mask $\mathbf{E}$. $\mathbf{E}(x, y) = 1$ represents that the pixel at position (x, y) belongs to the foreground, and $\mathbf{E}(x, y) = 0$ represents that the pixel belongs to the background.

In the following, we show how to use the generated foreground-background mask $\mathbf{E}$ to select local features from non-occluded areas and then fuse the global features and local features for matching pedestrian images.

We first define P evenly partitioned patches on mask $\mathbf{E}$ in the horizontal direction. Hence, the visibility score v_p ($p = 1, ..., P$) for each patch can be calculated as

$$v_p = \frac{1}{d' \times w'} \sum_{(x,y) \in \mathbf{\Omega}_p} \mathbf{E}(x, y), \tag{11}$$

where $d' = h'/P$ denotes the height for each patch, and $\mathbf{\Omega}_p$ is the p-th rectangular patch on mask $\mathbf{E}$.

Obviously, $v_p \in [0, 1]$ indicates the likelihood of foreground for the p-th patch. Thus, a large value of v_p indicates that the corresponding patch is likely to be non-occluded. In this way, the non-occlusion indicator I_p that infers the non-occluded areas can be calculated as

$$I_p = \Pi\{v_p \geq \tau\}, \tag{12}$$

where $I_p \in \{0, 1\}$ and τ is the threshold. The indicator function $\Pi\{\cdot\}$ takes on the value 1 if its argument is true, and 0 otherwise. When v_p is equal to or greater than the threshold τ, the p-th patch is considered to be non-occluded and can provide useful foreground information, and thus, I_p is set to 1. Otherwise, the corresponding patch is considered to be heavily occluded, and thus, I_p is set to 0.

In this paper, pedestrian matching is performed based on the local features extracted from the non-occluded areas according to the foreground-background mask from the semantic branch and the global features. The detailed pedestrian matching strategy is described as follows.

Assume that $\mathbf{Q}$ and $\mathbf{G}$ are the images from the query and the gallery, respectively. By using the trained SORN model, we can obtain the global features, local features and the corresponding non-occlusion indicators as $\{\mathbf{f}^Q_{global}, \{\mathbf{f}^{1,Q}_{local}, I^Q_1\}, ..., \{\mathbf{f}^{P,Q}_{local}, I^Q_P\}\}$ and $\{\mathbf{f}^G_{global}, \{\mathbf{f}^{1,G}_{local}, I^G_1\}, ..., \{\mathbf{f}^{P,G}_{local}, I^G_P\}\}$, of $\mathbf{Q}$ and $\mathbf{G}$ from the three branches, respectively.

The global distance of global features between $\mathbf{Q}$ and $\mathbf{G}$ can be calculated as

$$D^{QG}_{global} = \frac{\mathbf{f}^Q_{global} \cdot \mathbf{f}^G_{global}}{\|\mathbf{f}^Q_{global}\| \|\mathbf{f}^G_{global}\|}, \tag{13}$$

where $\|\cdot\|$ denotes the ℓ_2-norm.

The local distance of the p-th local features between $\mathbf{Q}$ and $\mathbf{G}$ can be calculated as

$$D^{QG}_{p} = \frac{\mathbf{f}^{p,Q}_{local} \cdot \mathbf{f}^{p,G}_{local}}{\|\mathbf{f}^{p,Q}_{local}\| \|\mathbf{f}^{p,G}_{local}\|}. \tag{14}$$

Therefore, the overall distance between $\mathbf{Q}$ and $\mathbf{G}$ is computed as

$$D^{QG}_{total} = \frac{D^{QG}_{global} + \sum_{p=1}^{P} I^Q_p I^G_p D^{QG}_p}{1 + \sum_{p=1}^{P} I^Q_p I^G_p}. \tag{15}$$

Based on Eq. (15), if a patch is heavily occluded in the query or gallery images, the local distance of this patch is not considered for computing the overall distance. Note that, even when there are no common non-occluded areas at the same patch locations (i.e., $I^Q_p \neq I^G_p$) for the query and gallery images, the global features that cover the full receptive field can still provide useful information for pedestrian matching.

In [12], PGFA also employs a similar pedestrian matching strategy, where the confidence scores of the detected landmarks are used to determine the occluded area. In contrast, we take advantage of the visibility score for each patch to indicate the occluded area, which makes our method insensitive to small segmentation errors.

We summarize the pedestrian matching strategy in Algorithm 1.

Algorithm 1 Pedestrian matching between a query image and a gallery image.

Input:
A query image $\mathbf{Q}$; A gallery image $\mathbf{G}$; The trained SORN model; The number of patches P; Threshold τ.

Output:
The distance D^{QG}_{total} between the query and gallery images.

1: Extracting the global features $\mathbf{f}^Q_{global}$, $\mathbf{f}^G_{global}$, local features $\{\mathbf{f}^{p,G}_{local}\}^P_{p=1}$, $\{\mathbf{f}^{p,Q}_{local}\}^P_{p=1}$ and the foreground-background masks $\mathbf{E}_Q$, $\mathbf{E}_G$ based on the trained SORN model for the query and gallery images, respectively.
2: Calculating the visibility scores$\{v^Q_p\}^P_{p=1}$ and $\{v^G_p\}^P_{p=1}$ via Eq. (11) for the query and gallery images, respectively.
3: Using the threshold τ to obtain the non-occlusion indicators $\{I^Q_p\}^P_{p=1}$ and $\{I^G_p\}^P_{p=1}$ via Eq. (12) for the query and gallery images, respectively.
4: Calculating the global distance D^{QG}_{global} of global features via Eq. (13).
5: Calculating the local distance $\{D^{QG}_p\}^P_{p=1}$ of local features via Eq. (14).
6: Calculating the overall distance D^{qg}_{total} of the query and gallery images via Eq. (15).

F. Discussions

It is worth noting that both our proposed method and some previous methods [8], [24], [25], [32] take advantage of semantic segmentation for person Re-ID. However, there are significant differences between our proposed method and these methods.

First, our proposed method makes use of semantic segmentation to specifically address the problem of occluded person Re-ID. In particular, our method employs the generated semantic masks from the semantic branch to indicate the non-occluded areas and select local features from these areas. In this way, the negative influence of occlusion can be greatly mitigated. In contrast, previous methods [8], [24], [32] make use of semantic segmentation (human parsing models are usually employed) to extract aligned local features and capture foreground areas. Therefore, these methods may not perform well on occluded person Re-ID, mainly because the occluded areas seriously affect the extraction of local features.

Second, our proposed method jointly trains the semantic branch with the other two branches in a multi-task learning manner. Semantic segmentation serves as an auxiliary task of occluded person Re-ID, and the joint training of two tasks (i.e., semantic segmentation and occluded person Re-ID) can help to effectively improve the performance of occluded person Re-ID. However, the human parsing models in previous methods [8], [24], [32] are only trained on external human semantic segmentation datasets since the ground-truth semantic labels are not available in person Re-ID datasets. In other words, the two tasks are separately trained. Hence, the intrinsic relationship between these two tasks is not taken into account. Some methods (such as [25]) employ semantic labels to guide the training of attention maps for extracting global and local features. Nevertheless, these methods are not applicable to the task of occluded person Re-ID, because the attention maps contain the disturbance information from the occluded areas.

IV. Experiments

In this section, extensive experiments are performed to evaluate the performance of the proposed SORN method. In Section IV-A, we introduce datasets and evaluation metrics. In Section IV-B, we provide the implementation details. In Section IV-C, we perform ablation studies to evaluate the key components of the proposed SORN method. In Section IV-D, we compare the proposed SORN method with several state-of-the-art person Re-ID methods. Finally, in Section IV-E, we show the parameter analysis.

A. Datasets and Evaluation Metrics

To show the effectiveness of the proposed SORN method, we perform experiments on an occluded person Re-ID dataset (Occluded-DukeMTMC) [12] and two partial person Re-ID datasets (Partial-REID [14] and Partial-iLIDS [13]). Moreover, we also evaluate the proposed method on two general person Re-ID datasets, including Market-1501 [43] and DukeMTMC-reID [44] [45], to further verify the generalization of the proposed method to deal with the problem of general person Re-ID.

Occluded-DukeMTMC is a large-scale occluded person Re-ID dataset collected from DukeMTMC-reID. The training set contains 15,618 images from 702 identities. The test set contains 2,210 query images and 17,661 gallery images from 1,110 identities. The training set, query set, and gallery set of Occluded-DukeMTMC contain 9%, 100%, and 10% occluded images in the respective sets. Therefore, all the query images are occluded, and the gallery set also contains a certain number of occluded images. Note that the training set and the test set of Occluded-DukeMTMC have different obstacles in case the trained model "remembers" the particular occlusion pattern in the inference stage.

Partial-REID and Partial-iLIDS are used to verify the effectiveness of our method on the problem of partial person Re-ID. Partial-REID contains 600 images from 60 identities. For each identity, there are five full-body images and five partial images. The images are collected at a university campus with different viewpoints, backgrounds, and different types of severe occlusion. All the query images are occluded, while all the gallery images are non-occluded. Partial-iLIDS is a partial person Re-ID dataset derived from iLIDS [46]. It contains 238 images from 119 identities collected in an airport, where the lower-body parts of pedestrians are often occluded by the luggage.

Market-1501 contains 1,501 identities captured by six cameras, where the dataset is split into 12,936 training images, 3,368 query images, and 19,732 gallery images. DukeMTMC-reID consists of 16,522 training images, 2,228 query images, and 17,661 gallery images from 1,404 identities.

We adopt the cumulative matching characteristic curves (CMC) and mean average precision (mAP) to evaluate the performance of person Re-ID methods. Here, R-1, R-5, and R-10 denote CMC at Rank-1, Rank-5, and Rank-10, respectively. All the following experiments are performed under the setting of a single query image.

B. Implementation Details

We use the PyTorch framework to implement the proposed SORN method, and a single GTX 2080 GPU is used for training and testing. The backbone network of our method is based on ResNet50 [38] (which is pretrained on ImageNet [47]), where the stride of the last residual block is set to 1. Following [48], we add a batch normalization layer [49] after the global average pooling layer in both the global and local branches.

The input image is resized to 384×128 for all experiments. For data augmentation, only random horizontal flip is used. The batch size is set to 32. We use the Adam optimizer [50] to minimize the loss. The total training takes 80 epochs. In the first 10 epochs, the learning rate is linearly increased from 3×10^{-5} to 3×10^{-4} and then decayed to 3×10^{-5} and 3×10^{-6} at the 30th epoch and the 60th epoch, respectively.

The regularization parameters λ_1, λ_2, and λ_3 are empirically set to 0.75, 0.50, and 1.50, respectively. The threshold τ is set to 0.15. In our implementation, the numbers of patches P and channel sets N are set to be the same value 3.

TABLE I
THE INFLUENCE OF GLOBAL AND LOCAL BRANCHES ON THE R-1, R-5, AND R-10 ACCURACY (%) AND MAP (%) ON THE OCCLUDED-DUKEMTMC DATASET. THE BEST RESULTS ARE BOLDFACED.

Method	R-1	R-5	R-10	mAP
Global+Semantic	48.8	63.6	69.8	38.1
Local+Semantic	48.0	61.8	66.2	39.0
SORN	**57.6**	**73.7**	**79.0**	**46.3**

TABLE II
THE INFLUENCE OF THE SPC LOSS AND PEDESTRIAN MATCHING STRATEGY ON THE R-1, R-5, AND R-10 ACCURACY (%) AND MAP (%) ON THE OCCLUDED-DUKEMTMC DATASET. THE BEST RESULTS ARE BOLDFACED.

Method		R-1	R-5	R-10	mAP
w/o SPC Loss	Global Features	51.5	66.5	71.9	41.6
	Global+Local Features	50.0	63.2	68.5	41.0
	SORN	55.3	71.4	77.1	44.3
SPC Loss	Global Features	54.1	69.7	74.3	44.1
	Global+Local Features	50.2	65.1	70.9	41.7
	SORN	**57.6**	**73.7**	**79.0**	**46.3**

C. Ablation Studies

We conduct experiments on the Occluded-DukeMTMC dataset to demonstrate the effectiveness of each component of our proposed method for occluded person Re-ID.

1) Effectiveness of the Local Branch and the Global Branch: In this subsection, we verify the effectiveness of the local branch and the global branch. We evaluate two variants of our proposed method: (1) the method (denoted as "Global+Semantic") that adopts the global branch and the semantic branch; and (2) the method (denoted as "Local+Semantic") that uses the local branch and the semantic branch. We also evaluate the proposed SORN method, which effectively combines the global, semantic, and local branches in an integrated network. Table I gives the results obtained by two variants and the proposed method.

From Table I, we can see that our proposed SORN method significantly outperforms the two variants. This is mainly because the local branch learns fine-grained local features, while the global branch learns salient global appearance representations with occlusion awareness. Therefore, these two branches can complement each other, leading to the improvement of the final performance.

2) Effectiveness of the SPC Loss and Pedestrian Matching: In this subsection, we evaluate the effectiveness of the proposed SPC loss and the proposed pedestrian matching strategy. We evaluate two variants of our proposed method that use different pedestrian matching strategies: (1) the method (denoted as "Global Features") that only uses the global features for matching the query image and the gallery images; and (2) the method (denoted as "Global+Local Features") that uses the global features and all local features extracted from the whole pedestrian image for matching. The proposed SORN method that employs the proposed pedestrian matching strategy is also used for a comparison. The three methods are evaluated by their models trained with the SPC loss and without the SPC loss (denoted as "w/o SPC Loss"). Table II gives the performance comparison obtained by these methods.

In Table II, we can see that the three methods trained with the SPC loss achieve much better performance than those trained without the SPC loss. Specifically, the Rank-1 and mAP obtained by the Global Features method trained with the SPC loss are improved by 2.6% and 2.5%, respectively, compared with that trained without the SPC loss. The Rank-1 and mAP obtained by the proposed SORN method are respectively increased by 2.3% and 2.0% in comparison with that trained without the SPC loss. The above results verify the importance of the proposed SPC loss. The SPC loss effectively enforces the model to learn occlusion-robust global features, which are beneficial for improving the final performance.

Note that the Global Features method obtains better performance (in terms of both CMC and mAP) than the Global+Local Features method. This is mainly because the local features extracted from the local branch contain noise due to the occlusion disturbance, thus leading to a performance decrease. However, compared with the Global Features method, the proposed SORN achieves better performance. In particular, the Rank-1 accuracy and mAP obtained by the SORN trained with the SPC loss are respectively increased by 3.5% and 2.2%, compared with the Global Features method trained with the SPC loss. This demonstrates the effectiveness of the proposed pedestrian matching strategy.

3) Effectiveness of the Semantic Branch: In this subsection, we verify the effectiveness of the semantic branch. We evaluate several methods (including the Global Features method, the Global+Local Features method and the proposed SORN method) under different settings. "w/o Semantic" denotes that we train the model without using the semantic branch. "2 Labels" denotes that 2 semantic labels (i.e., foreground and background) are predicted in the semantic branch. "5 Labels" denotes that 5 semantic labels (i.e., head, torso, arm, leg, and background) are predicted in the semantic branch. "8 Labels" denotes that all 8 semantic labels are predicted in the semantic branch. Note that "Global+Local Features+S" under the setting of "w/o Semantic" denotes that we train an external semantic segmentation model (for a fair comparison, we use an external semantic segmentation model having the same network architecture as our method and use all 8 semantic labels to train the network) and employ it to select non-occluded local features. Furthermore, "Global+Local Features+P" under the setting of "w/o Semantic" denotes that we employ an external pose estimation model (the AlphaPose model [51] is used) to select non-occluded local features. Table III shows the performance comparison obtained by these methods.

Compared with the Global+Local Features+S method, SORN with 8 labels performs better by 1.8% in terms of Rank-1 accuracy and 3.4% in terms of mAP. Therefore, jointly training the tasks of occluded person Re-ID and semantic segmentation in a multi-task learning framework can effectively improve the performance. In other words, the auxiliary task of semantic segmentation is beneficial for increasing the performance of occluded person Re-ID. Meanwhile, SORN with 8 labels outperforms the Global+Local Features method with 8 labels by 7.4% in terms of Rank-1 accuracy and 4.6% in terms of mAP. This verifies the importance of removing the disturbance of occluded areas during pedestrian matching.

According to the above results, the use of the semantic branch to alleviate the negative effect of occluded areas and background clutter provides more performance improvements than the joint training of two tasks.

For the Global+Local Features+S method and the Global+Local Features+P method, the semantic information (8 semantic labels) and the landmark information (18 landmarks) provided by externally trained models are respectively used. Therefore, the two methods achieve better results than the Global Features and Global+Local Features methods using 2 labels since more detailed semantic information or landmark information is employed. Meanwhile, the Global Features and Global+Local Features methods using 2 labels achieve similar results to those without using the semantic branch. The main reason is that coarse semantic labeling cannot provide sufficient information for extracting effective local features to deal with occlusion when only 2 labels are predicted in the semantic branch.

It is worth noting that the performance of the Global+Local Features+P method (using an external pose estimation model) is worse than that of the Global+Local Features+S method (using an external semantic segmentation model). This is because the occlusion information obtained by the semantic segmentation model is more fine-grained and robust than that obtained by the pose estimation model. The pose estimation model only predicts a small number of keypoints. However, a patch (e.g., a partially occluded patch) may still contain useful information, even though no keypoint is detected in this region.

In addition, the performance of the proposed SORN is improved, when more semantic labels are used. In particular, SORN using all 8 labels in the semantic branch respectively improves Rank-1 accuracy and mAP by 2.8% and 4.4%, compared with that using only 2 labels. This indicates that coarse semantic labeling is not advantageous for improving the final performance. In other words, the semantic branch trained with more fine-grained semantic labels can help to enhance the performance of our method for occluded person Re-ID. This is due to two reasons: 1) fine-grained body-part labels provide detailed position information, which is useful for alleviating the misalignment problem; and 2) fine-grained labels can make the prediction of the semantic branch more accurate, and these labels can be used to determine whether a particular patch is occluded.

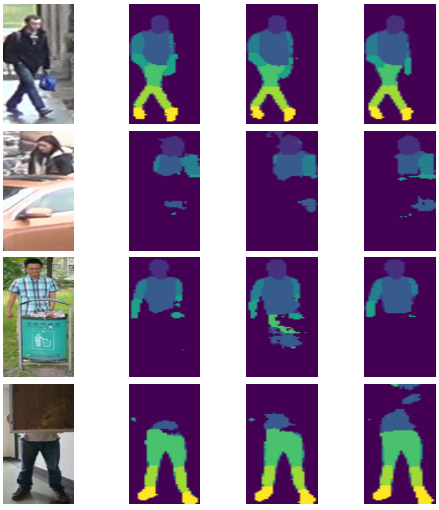
4) Effectiveness of Label Smoothing: In this subsection, we evaluate the effectiveness of label smoothing. We qualitatively give the segmentation masks of some images in the Occluded-DukeMTMC dataset, as shown in Fig. 5. Specifically, we show the segmentation masks obtained by our SORN (with label smoothing), SORN without label smoothing (denoted as SORN w/o LS), and an external model. For a fair comparison, the external model adopts the same architecture as the semantic branch of the SORN and uses label smoothing for training. For all the methods, 8 semantic labels are used to train the models.

For the non-occluded pedestrian image (see the first row of Fig. 5), the segmentation results obtained by the competing methods are similar. However, for the occluded pedestrian images (see the second row to the fourth row of Fig. 5), the segmentation results obtained by the SORN are better than those obtained by the SORN w/o LS. In particular, the segmentation masks generated by the SORN contain less noise in occluded areas. Therefore, label smoothing helps improve the segmentation results since it can prevent overfitting on the training data containing false semantic labels. In this way, the masks generated by the SORN can be used to effectively select non-occluded local features. Compared with the external model, the SORN still achieves better segmentation results. Hence, the joint training of occluded person Re-ID and semantic segmentation improves the performance of each task.

Moreover, we also evaluate the influence of label smoothing on the performance in Table III. The SORN achieves better results than the SORN w/o LS when 5 or 8 labels are used. Label smoothing plays an important role in improving the performance of occluded person Re-ID. However, the SORN obtains worse results than the SORN w/o LS when 2 labels

TABLE III
THE INFLUENCE OF THE SEMANTIC BRANCH ON THE R-1, R-5, AND R-10 ACCURACY (%) AND MAP (%) ON THE OCCLUDED-DUKEMTMC DATASET. THE BEST RESULTS ARE BOLDFACED.

Method		R-1	R-5	R-10	mAP
w/o Semantic	Global Features	53.3	67.5	73.0	41.4
	Global+Local Features	49.8	64.1	69.7	39.7
	Global+Local Features+S	55.8	71.4	76.7	42.9
	Global+Local Features+P	55.2	70.4	76.6	42.2
2 Labels	Global Features	52.8	67.5	74.0	41.2
	Global+Local Features	49.5	64.4	70.3	39.5
	SORN w/o LS	55.1	70.2	76.6	42.1
	SORN	54.8	69.9	75.8	41.9
5 Labels	Global Features	53.9	67.9	72.9	42.6
	Global+Local Features	50.4	64.2	68.8	40.6
	SORN w/o LS	55.3	71.8	77.4	44.1
	SORN	56.3	72.5	77.7	44.5
8 Labels	Global Features	54.1	69.7	74.3	44.1
	Global+Local Features	50.2	65.1	70.9	41.7
	SORN w/o LS	55.6	71.8	77.9	44.8
	SORN	**57.6**	**73.7**	**79.0**	**46.3**



(a) (b) (c) (d)

Fig. 5. Examples of generated segmentation masks. From left to right, (a) the input images, the segmentation masks obtained by (b) SORN, (c) SORN w/o LS, and (d) an external model.

are used. This shows the negative influence of label smoothing on the performance in the case of coarse semantic labeling.

D. Comparisons with State-of-the-Art Methods

In this section, we compare our proposed SORN method with several state-of-the-art methods on the Occluded-DukeMTMC, Partial-REID, Partial-iLIDS, Market-1501, and DukeMTMC-reID datasets.

1) Results on Occluded-DukeMTMC: Table IV shows the performance obtained by our proposed SORN method and twelve competing methods, including LOMO+XQDA [52], DIM [53], Part Aligned [54], Random Erasing [55], HACNN [56], AOS [23], PCB [3], Part Bilinear [31], FD-GAN [9], DSR [15], SFR [16], PGFA [12], and HOReID [36], on the Occluded-DukeMTMC dataset. Note that TCSDO [37] requires an extra training dataset to train the teacher network and FPR [35] assumes the gallery images are non-occluded. Therefore, TCSDO and FPR are not used for comparisons.

LOMO+XQDA, DIM, Part Aligned, Random Erasing, HACNN, AOS, and PCB are the representative methods proposed for general person Re-ID. The proposed SORN method achieves 57.6% Rank-1 accuracy and 46.3% mAP on the Occluded-DukeMTMC dataset and outperforms these methods by a large margin. Compared with Part Bilinear [31] and FD-GAN [9] that make use of the external pose estimation models, our method achieves better performance. Our method integrates a semantic branch into a multi-task learning network and does not rely on externally trained models. DSR [15] and SFR [16] are specifically designed for the problem of partial person Re-ID. However, our method still obtains better performance than these methods. This is because our method takes advantage of semantic segmentation to deal with occlusion. PGFA [12] and HOReID [36] are state-of-the-art occluded person Re-ID methods, and our method significantly outperforms them by 6.2%/2.5% in terms of Rank-1 accuracy and 9.0%/2.5% in terms of mAP, which demonstrates the superiority of the proposed method.

Fig. 6 shows two very challenging query images and the corresponding top 5 retrieval results obtained by our proposed method and the Global+Local Features method (without using the semantic branch). The Global+Local Features method fails to accurately retrieve the correct results when the query images are severely occluded. This can be ascribed to the fact that the extracted global and local features are contaminated with the noise from the occluded areas. In contrast, our proposed method achieves promising results. This demonstrates the effectiveness of the proposed method in dealing with the problem of occluded person Re-ID.

2) Results on Partial-REID and Partial-iLIDS: Table V shows the performance obtained by our method and several competing methods, including MTRC [57], AMC+SWM [14], DSR [15], SFR [16], VPM [18], PGFA [12], FPR [35], and HOReID [36], on the Partial-REID and Partial-iLIDS datasets. Following the same settings as previous methods [12], [18], our method is trained on Market-1501 and tested on these two partial Re-ID datasets.

The proposed method achieves 76.7% and 79.8% Rank-1 accuracy on Partial-REID and Partial-iLIDS, respectively.

TABLE IV
PERFORMANCE COMPARISON IN TERMS OF R-1, R-5, AND R-10 ACCURACY (%) AND MAP (%) ON THE OCCLUDED-DUKEMTMC DATASET. THE BEST RESULTS ARE BOLDFACED.

Method	R-1	R-5	R-10	mAP
LOMO+XQDA [52]	8.1	17.0	22.0	5.0
DIM [53]	21.5	36.1	42.8	14.4
Part Aligned [54]	28.8	44.6	51.0	20.2
Random Erasing [55]	40.5	59.6	66.8	30.0
HACNN [56]	34.4	51.9	59.4	26.0
AOS [23]	44.5	-	-	32.2
PCB [3]	42.6	57.1	62.9	33.7
Part Bilinear [31]	36.9	-	-	-
FD-GAN [9]	40.8	-	-	-
DSR [15]	40.8	58.2	65.2	30.4
SFR [16]	42.3	60.3	67.3	32.0
PGFA [12]	51.4	68.6	74.9	37.3
HOReID [36]	55.1	-	-	43.8
SORN (ours)	**57.6**	**73.7**	**79.0**	**46.3**

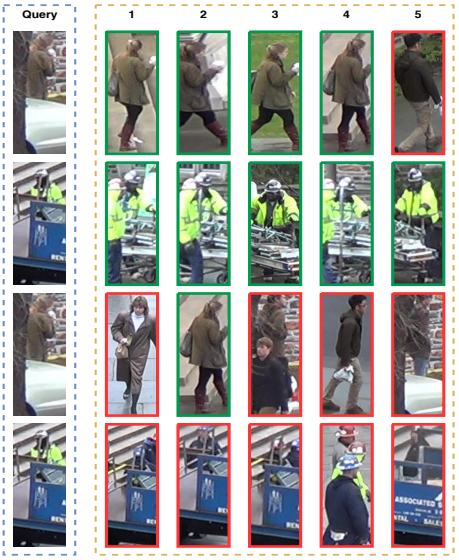


Fig. 6. The top 5 retrieval results obtained by the proposed SORN method (the first two rows) and the Global+Local Features method (the last two rows) for two query images in the Occluded-DukeMTMC Re-ID dataset. The images with green and red borders denote correct and incorrect results, respectively.

Note that MTRC, AMC+SWM, DSR, SFR, STNReID, and VPM are developed for partial person Re-ID, and they need to manually crop the occluded areas of query images. In contrast, our method can directly use the occluded images as the query. MTRC [57] and AMC+SWM [14] are based on handcrafted features. The proposed method outperforms them by a large margin, which shows the excellent advantage of deep neural networks. VPM [18] utilizes self-supervision to obtain visibility scores and select local features. However, the manual cropping process may discard some useful information. Hence, the performance of VPM is inferior to that of our method. Moreover, our method outperforms PGFA by 6.3% and 10.7% improvements in terms of Rank-1 accuracy on the Partial-REID and Partial-iLIDS datasets, respectively. This demonstrates the effectiveness of the proposed method for partial person Re-ID.

Compared with the recently proposed occluded person Re-ID methods (such as FPR and HOReID), the SORN achieves

TABLE V
PERFORMANCE COMPARISON IN TERMS OF R-1 AND R-3 ACCURACY (%) AND MAP (%) ON PARTIAL-REID AND PARTIAL-ILIDS. THE BEST RESULTS ARE BOLDFACED.

Method	Partial-REID		Partial-iLIDS	
	R-1	R-3	R-1	R-3
MTRC [57]	23.7	27.3	17.7	26.1
AMC+SWM [14]	37.3	46.0	21.0	32.8
DSR [15]	50.7	70.0	58.8	67.2
SFR [16]	56.9	78.5	63.9	74.8
STNReID [17]	66.7	80.3	54.6	76.3
VPM [18]	67.7	81.9	67.2	76.5
PGFA [12]	68.0	80.0	69.1	80.9
FPR [35]	81.0	-	68.1	-
HOReID [36]	**85.3**	**91.0**	72.6	86.4
SORN (ours)	76.7	84.3	**79.8**	**86.6**

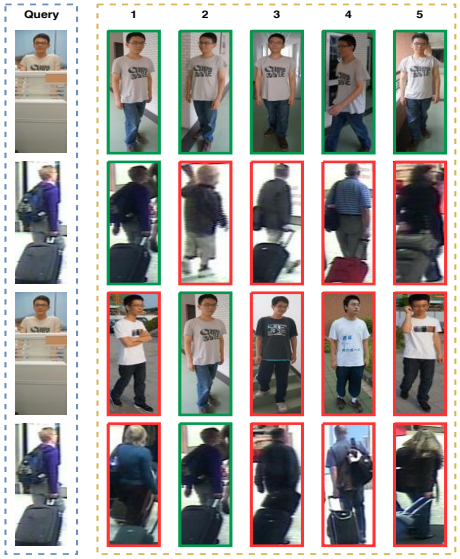


Fig. 7. The top 5 retrieval results obtained by the proposed SORN method (the first two rows) and the Global+Local Features method (the last two rows) for one query image in the Partial-REID dataset and one query image in the Partial-iLIDS dataset. The images with green and red borders denote correct and incorrect results, respectively.

better performance on the Partial-iLIDS dataset. However, the SORN is inferior to these two methods on the Partial-REID dataset. This is mainly because we train our model on Market-1501 since Partial-REID does not provide a training dataset. The domain gap between Partial-REID and Market-1501 is large, which affects the prediction accuracy of semantic labels and thus reduces the performance of SORN.

Fig. 7 shows two challenging query images and the corresponding top 5 retrieval results obtained by our proposed method and the Global+Local Features method (without using the semantic branch) on the Partial-REID and the Partial-iLIDS datasets. The Global+Local Features method obtains worse retrieval results than our proposed method. Note that there exists only one image having the same identity as the query image in the gallery set of Partial-iLIDS. Therefore, the second to fifth retrieval results in the second row of Fig. 7 cannot be correct, when the top retrieval 1 result is correct.

3) Results on Market-1501 and DukeMTMC-reID: Table VI shows the performance evaluation obtained by our method

TABLE VI
PERFORMANCE COMPARISON IN TERMS OF R-1 ACCURACY (%) AND MAP (%) ON MARKET-1501 AND DUKEMTMC-REID. THE BEST RESULTS ARE BOLDFACED.

Method	Market-1501		Duke-MTMC-reID	
	R-1	mAP	R-1	mAP
BoW+Kissme [43]	44.4	20.8	25.1	12.2
SVDNet [2]	82.3	62.1	76.7	56.8
PAN [45]	82.8	63.4	71.7	51.5
PAR [54]	81.0	63.4	-	-
PAN [20]	82.0	63.0	-	-
DSR [15]	83.5	64.2	-	-
Triplet loss [27]	84.9	69.1	-	-
Quadruplet loss [1]	86.3	72.2	73.4	58.0
AOS [23]	86.5	78.3	79.1	62.1
APR [58]	87.0	66.9	73.9	55.6
DPFL [59]	88.9	73.1	79.2	60.6
MLFN [60]	90.0	74.3	81.0	62.8
PCB [3]	92.4	77.3	81.9	65.3
SPReID [8]	92.3	80.5	83.8	69.3
PGFA [12]	91.2	76.8	82.6	65.5
SORN (ours)	**94.8**	**84.5**	**86.9**	**74.1**

and other competing methods, including BoW+Kissme [43], SVDNet [2], PAN [45], PAR [54], Pedestrian [20], DSR [15], Triplet loss [27], Quadruplet loss [1], AOS [23], APR [58], DPFL [59], MLFN [60], PCB [3], SPReID [8], and PGFA [12], on Market-1501 and DukeMTMC-ReID.

The proposed method, which effectively combines the local features from the non-occluded areas and the occlusion-robust global features, outperforms the methods based on local features (including PAN [45], PAR [54], and PCB [3]) and the methods based on multi-level and multi-scale features (including DPFL [59] and MLFN [60]). Note that our method uses the simple classification loss and the SPC loss, but it still obtains better performance than the methods based on the specifically designed losses, such as Triplet loss [27] and Quadruplet loss [1]. Moreover, our method outperforms SPReID [8], which also uses semantic segmentation for person Re-ID. In summary, our method achieves state-of-the-art performance for general person Re-ID.

E. Parameter Analysis

In this section, we conduct experiments to evaluate the influence of several key parameters in the proposed method (including the threshold τ in Eq. (12), the number of patches P, and the number of channel sets N) on the performance (in terms of Rank-1 accuracy and mAP). We use the Occluded-DukeMTMC and Partial-REID datasets for parameter analysis. Here, we change the values of one parameter and fix the other parameters for analysis.

1) Influence of the Threshold τ: τ is the threshold used to select the non-occluded patches. Fig. 8(a) shows the influence of threshold τ on the Rank-1 accuracy and mAP. When τ is small, some severely occluded patches are used, and thus the performance is poor. When τ is large, the patches with slight occlusion are discarded, but these patches may contain useful information. Therefore, the performance decreases. When $\tau = 0.15$, our method achieves the best performance.

2) Influence of the Number of Patches P: Fig. 8(b) shows the influence of the number of patches P on the final performance. Our method is sensitive to the number of patches P in

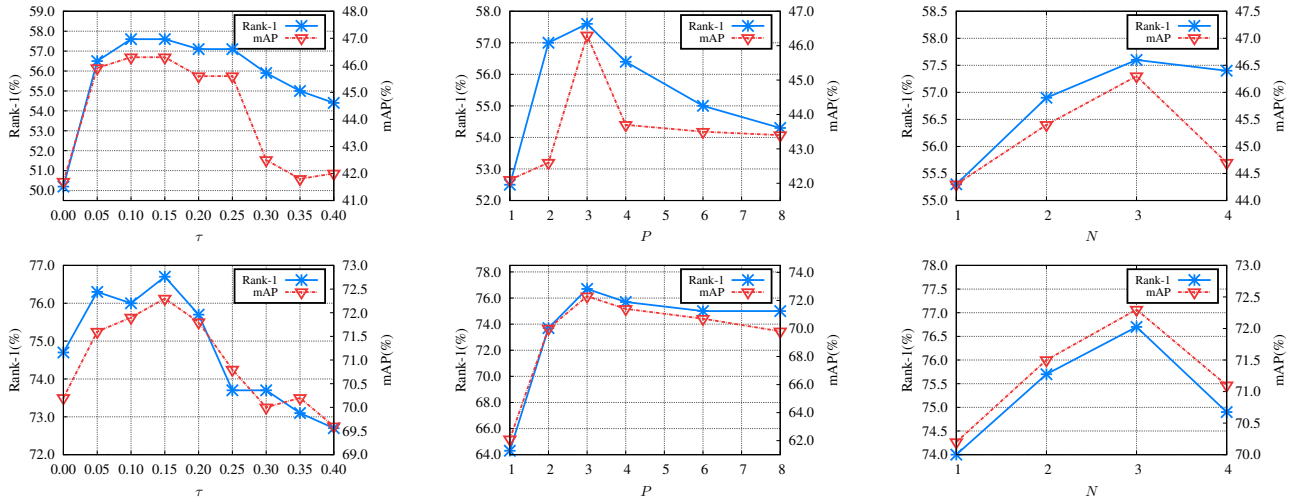


Fig. 8. The Rank-1 accuracy and mAP obtained by the proposed method with the different values of (a) the threshold τ, (b) the number of patches P, and (c) the number of channel sets N on the Occluded-DukeMTMC dataset (the first row) and the Partial-REID dataset (the second row).

terms of Rank-1 and mAP. In particular, our method obtains the top performance on the two datasets when the number of patches is set to 3. This shows that P can be empirically chosen to a fixed value. On the one hand, when the number of patches increases, the patch size becomes larger. As a result, the local features extracted from the large patches cannot effectively provide discriminative local information. On the other hand, when the number of patches is too large, the patch size is small. In this case, the semantically consistent human body is over-segmented into many small patches. Hence, the local context cannot be effectively captured from the small patches, and thus the performance degenerates.

3) Influence of the Number of Channel Sets N: Fig. 8(c) shows the influence of the number of channel sets N on the final performance. Specifically, we fix the number of patches P to 3 and change N from 1 to 4. When the value of N is set to 3, our method obtains the best performance on the two datasets. Hence, N can be empirically set to 3 to generally ensure the retrieval performance of our method. Note that each normalized aggregation map is divided into N horizontal patches. Therefore, it is beneficial to extract discriminative global features at a moderate patch size since the sizes and positions of obstacles vary. Too large or too small values of N adversely affect the extraction of effective global features, thus leading to a performance decrease. When the value of N is equal to 1, our method is trained without using the SPC loss, and it achieves the worst performance.

V. Conclusion

In this paper, we propose a novel SORN method, a three-branch (consisting of the global, local, and semantic branches) CNN, for occluded person Re-ID. For the global branch, we develop a new SPC loss, which enables the extracted global features to encode occlusion-aware local information and thus makes the extracted features robust to occlusion. For the local branch, the fine-grained local features are extracted. For the semantic branch, the semantic mask is generated to indicate the non-occluded areas. These three branches are jointly trained in a multi-task learning framework. Extensive experiments show the effectiveness of SORN for the problems of occluded, partial, and general person Re-ID.

References

[1] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 403–412.

[2] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3800–3808.

[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. Comput. Vis. (ECCV)*, Sep. 2018, pp. 480–496.

[4] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis. Comput. Vis. (ECCV)*, Oct. 2016, pp.475–491.

[5] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

[6] 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. (ICCV)*, Oct. 2017, pp. 3960–3969.

[7] J. Lei, L. Niu, H. Fu, B. Peng, Q. Huang, and C. Hou, "Person re-identification by semantic region representation and topology constrain-t," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2453–2466, 2019.

[8] 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. (CVPR)*, Jun. 2018, pp.1062–1071.

[9] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2018, pp.1222–1233.

[10] 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. (CVPR)*, Jun. 2017, pp. 384–393.

[11] 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.

[12] 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. (ICCV)*, Oct. 2019, pp. 542–551.

[13] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 649–656.

[14] 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. (ICCV)*, Dec. 2015, pp. 4678–4686.

[15] 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. (CVPR)*, Jun. 2018, pp. 7073–7082.

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

[17] 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.*, 2020.

[18] 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. (CVPR)*, Jun. 2019, pp. 393–402.

[19] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, Oct. 2016.

[20] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, 2019.

[21] C. Shen, G.-J. Qi, R. Jiang, Z. Jin, H. Yong, Y. Chen, and X.-S. Hua. "Sharp attention network via adaptive sampling for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3016–3027, 2019.

[22] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.

[23] 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. (CVPR)*, Jun. 2018, pp. 5098–5107.

[24] 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*, Apr. 2018.

[25] H. Cai, Z. Wang and J. Cheng, "Multi-scale body-part mask guided attention for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit Workshop. (CVPRW)*, Jun. 2019, pp. 1555-1564.

[26] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2014, pp. 34–39.

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

[28] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proc. Eur. Conf. Comput. Vis. Comput. Vis. (ECCV)*, Sep. 2018, pp. 188–204.

[29] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8649–8658.

[30] H. Hu, W. Fang, B. Li, and Q. Tian, "An adaptive multi-projection metric learning for person re-identification across non-overlapping cameras," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2809–2821, 2019.

[31] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis. Comput. Vis. (ECCV)*, Sep. 2018, pp. 402–419.

[32] 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. (CVPR)*, Jun. 2018, pp. 1179–1188.

[33] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[34] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. IEEE Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[35] 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," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2019, pp. 8449-8458.

[36] 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. (CVPR)*, Jun. 2020.

[37] J. Zhuo, J. Lai, and P. Chen, "A novel teacher-student learning framework for occluded person re-identification," *arXiv preprint arXiv:1907.03253*, Jul. 2019.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[39] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. Zhou, "Towards good practices on building effective cnn baseline model for person re-identification," *arXiv preprint arXiv:1807.11042*, Jul. 2018.

[40] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, 2017.

[41] R. Alp Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7297–7306.

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[43] 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. (ICCV)*, Oct. 2015, pp. 1116–1124.

[44] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis. Comput. Vis. (ECCV)*, Oct. 2016, pp. 17–35.

[45] 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. (ICCV)*, Oct. 2017, pp. 3754–3762.

[46] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. Eur. Conf. Comput. Vis. Comput. Vis. (ECCV)*, Sep. 2014, pp. 688–703.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[48] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. T. Zhou, "Towards good practices on building effective CNN baseline model for person re-identification," *arXiv preprint arXiv:1807.11042*, Jul. 2018.

[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, Mar. 2015.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014.

[51] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[52] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp.2197–2206.

[53] 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*, Nov. 2017.

[54] 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. (ICCV)*, Oct. 2017, pp. 3219–3228.

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

[56] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2285–2294.

[57] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. March. Intell.*, vol. 35, no. 5, pp. 1193–1205, 2013.

[58] 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 Recognit.*, vol. 95, pp. 151–161, 2019.

[59] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2590–2600.

[60] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2109–2118.