This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

Occlusion-sensitive Person Re-identification via Attribute-based Shift Attention

Hanyang Jin, Shenqi Lai, and Xueming Qian, *Member, IEEE*

***Abstract*—Occluded person re-identification is one of the most challenging tasks in security surveillance. Most existing methods focus on extracting human body features from occluded pedestrian images. This paper prioritizes a difference between occluded and non-occluded person re-ID: When computing the similarity between a holistic pedestrian image and an occluded pedestrian image, a certain part of the human body in this holistic image can be distractive for pedestrian retrieval. To solve this problem, we propose an occluded person re-ID framework named attribute-based shift attention network (ASAN). First, unlike other methods that use off-the-shelf tools to locate pedestrian body parts in the occluded images, we design an attribute-guided occlusion-sensitive pedestrian segmentation (AOPS) module. AOPS is a weakly supervised method that leverages the semantic-level attribute annotations in person re-ID datasets. Second, guided by the pedestrian masks provided by AOPS, a shift feature adaption (SFA) module extracts the visible part of the human body feature in a part-based manner. After that, a visible region matching (VRM) algorithm is proposed to filter out the interference information in the holistic person images during the retrieval phase and further purify the representation of pedestrian features. Extensive experiments with ablation analysis demonstrate our method's effectiveness. And the state-of-the-art results are achieved on four occluded datasets Partial-REID, Partial-iLIDS, Occluded-DukeMTMC, and Occluded REID. Moreover, the experiments on two holistic person re-ID datasets Market-1501 and DukeMTMC-reID, and a vehicle re-ID dataset VeRi-776 show that ASAN also has a good generality.**

***Index Terms*— Person Re-ID, Weakly Supervision, Feature Extraction, Attention Mechanism.**

I. Introduction

PERSON re-identification (re-ID) aims to search a probe (or query) pedestrian from dis-joint camera views. It is an important research topic in computer vision with many applications, such as unmanned driving, security monitoring, and behavior analysis. The mainstream models have achieved satisfactory performance on the public datasets [1]-[2], which generally either utilize global pedestrian features [3] or elaborately merged it with local features [44]-[47],[60]. However, these methods are designed with the assumption that the complete human body is visible. But in real scenes, the pedestrians are always inevitably obscured by trees, cars, walls, or other people. Therefore, when tested on occluded benchmarks, the traditional person re-ID methods often mistake the occlusion as a part of the human body and cannot perform well.

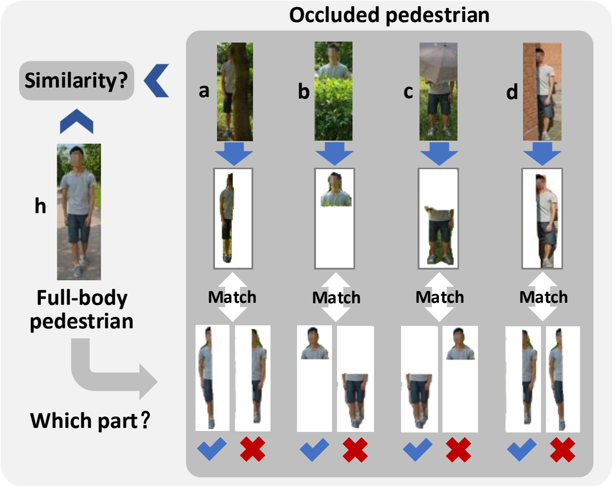

Fig. 1. When comparing a holistic pedestrian image with an occluded pedestrian image, a certain region of the human body in this holistic image can be distractive. And when compared with occluded images with different occlusion positions, the position of this region is also different.

The main challenge in the occluded person re-ID task is that features extracted by traditional models may involve not only the target person but also the occluded regions, which can easily corrupt the representation. Moreover, part-based local descriptions [44]-[47][55] have been proven to be robust and effective for holistic person re-ID, but strict body parts alignment is required in these methods so that they can hardly work well in the case of occlusion. Recently, several works [5],[7]-[13],[20],[33], [43] are proposed to solve the occlusion problem in person re-ID. In the setting of these works, occluded images are formed as the query to search the full-body images with the same identities in the gallery. And the strategy shared by these methods is to train a model that can extract the features of non-occluded parts from occluded pedestrian images.

However, on the one hand, the best performing occluded person re-ID methods [5],[20], [33], [43] use off-the-shelf tools like instance segmentation, key-points, or pose estimation models to locate the visible human body parts in occluded person images, which need extra annotation beyond person re-

This work was supported in part by the NSFC under Grant 61732008, 61772407.

Xueming Qian (corresponding author, qianxm@mail.xjtu.edu.cn) is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security and with SMILES LAB, Xi'an Jiaotong University, 710049, Xi'an, China.

Hanyang Jin (E-mail: jhy0606@stu.xjtu.edu.cn) and Shenqi Lai (E-mail: laishenqi@stu.xjtu.edu.cn) are with the SMILES LAB at School of Information and Communications Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

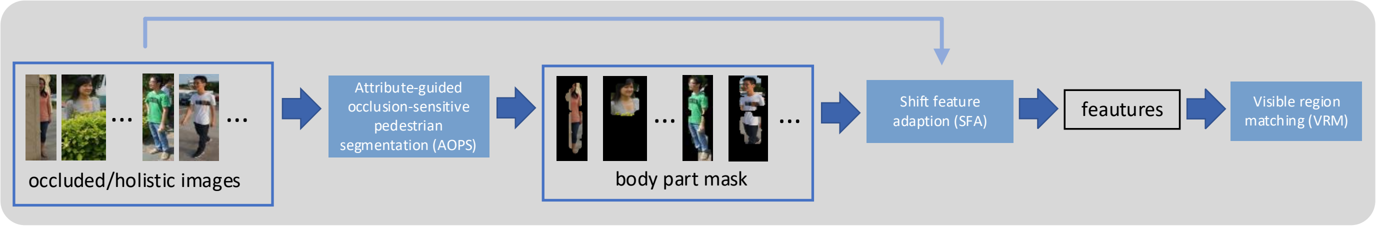
occluded/holistic images
Attribute-guided occlusion-sensitive pedestrian segmentation (AOPS)
body part mask
Shift feature adaption (SFA)
feautures
Visible region matching (VRM)

Fig.2 General pipeline of proposed framework. First, the occluded/holistic (in training, there are only holistic training samples, and in testing, there are both occluded queries and holistic galleries) images are fed into the AOPS module, the body part masks are obtained. Then, the images and masks are both input to the SFA, which learns a discriminative feature representation with a proposed occlusion batch painting strategy. Finally, the features are used to perform the VRM.

ID datasets and labor-intensive **pixel-level** annotations.

On the other hand, apart from extracting non-occluded human body features, there is another issue that should be specifically noticed in occluded person re-ID. As is shown in Fig. 1, when a full-body person image (h) and an occluded person image (one from $\{a,b,c,d\}$) are paired to compute the similarity, a certain region of this holistic image should emphatically be perceived by the feature extractor. But when compared with occluded images with different occlusion positions, this region should be **different**. For instance, when h is matched with a (the right half of the person is blocked), the left half region of the human body in h should be used to calculate the similarity with the left part of the pedestrian in a. While the upper half of h should be used when the comparison object is b (the lower half is occluded), because the lower body in b is not visible. This means that, although the goal of person re-ID on holistic datasets is to fully excavate the information in a pedestrian image, for occluded person re-ID, some part of a human body may bring redundant information and compromise the retrieval accuracy.

In order to tackle the problem mentioned above, this paper proposes an attribute-based shift attention network (ASAN) for occluded person re-ID.

First, we propose a weakly supervised pedestrian segmentation module for our occluded re-ID task, named attribute-guided occlusion-sensitive pedestrian segmentation (AOPS). With a novel interference occlusion erasing strategy in training, AOPS leverages attribute information and learns an occlusion-sensitive segmentation facility under weak supervision. Then, the attribute-based body part masks provided by AOPS are used to guide the shift feature adaption (SFA) module to generate the discriminative pedestrian features. In the testing end, a visible region matching (VRM) algorithm is performed to eliminate the interference of disturbing parts in holistic person images.

The major contribution of this work can be summarized as follows.

(1) To resist the obstruction in the occluded pedestrian images, instead of utilizing off-the-shelf tools to locate the body part, we manage to train a weakly supervised AOPS with the semantic-level attribute annotations of person re-ID datasets.

(2) To tackle the problem of interference caused by useless parts of the full-body picture, we first propose the SFA, which effectively extracts the visible body feature in a part-based manner, and then present the VRM to further purify the feature representation.

(3) Experiments on four authoritative occluded person re-ID benchmarks demonstrate the superiority of our method. And the experimental results on two holistic person re-ID datasets and a vehicle re-ID dataset also confirm that our method has good generalization ability.

The remainder of this paper is organized as follows: in section II, we review the related works. Then, following the pipeline described in Fig. 2, we first show the technical details of AOPS and SFA in section III and section IV. Then in section V, the loss function and the VRM are introduced. Experiments and analyses are shown in section VI. Finally, we conclude our work and discuss the future plan in Section VII.

II. Related Work

In this section, we briefly introduce existing works of three aspects: class activation map (CAM), pedestrian attribute leaning, and person re-ID, as they present to be preliminary knowledge of our work.

A. *Class Activation Map and Pedestrian Attribute learning*

This paper uses CAMs of pedestrian attributes to complete the task of localizing the pedestrian's body in the occluded pictures. So the related work of CAM and pedestrian attribute learning are discussed in this part.

Zhou *et al.* [4] proposed class activation map (CAM) for CNNs that uses average pooling layer and FC layer to modify the high layers of the classification network. When a classification network is trained, by locating the parameters of the fully connected layer corresponding to the category with a high predicted probability value and weighting them with the feature map, the image region with high response to this category can be highlighted [4]. Further, these regions can be used to complete rough semantic segmentation [14]-[15]. Unlike the training process of the supervised segmentation model, in the training of CAM and its derivatives [4], [14]-[15], the object location label (pixel-level annotation) is not given, but only the category label (image-level annotation). Thus, the supervision information is weaker, and the training of these methods is called weakly supervised learning.

Attributes are usually viewed as a mid-level semantic description for feature representation learning and have been investigated in numerous works. In [34], the authors applied several metric learning methods for the attribute's study. Su *et al.* [36] considered multiple cameras as related tasks and learned a discriminative network by multi-task learning. Khamis *et al.* jointly optimized the triplet loss for re-ID and attribute identity loss in [37]. In [38], fine-tuned CNN was embedded for attribute classification. Research [56] proposes a

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

self-supervised learning algorithm that is based on attribute-identity embedding. In addition, there are works [17][39]-[41] with regard to the specific datasets for attribute learning tasks. Thereinto, Lin *et al*. [17] manually annotated the Market-1501 [22] and DukeMTMC-reID [23] datasets with attribute labels. For every ID, there are adequate number of training samples for attribute learning, which the other attribute datasets don't have. Although some research [6][17] try to leverage attribute learning to improve the person search, pedestrian attribute recognition, which aims to predict the presence of a set of attributes (age groups, clothes types and whether holding bags, *etc.*) from an image, still remains a relatively independent task from person re-ID and hardly offer a substantial help for accurate person re-ID.

B. Occluded Person Re-ID

Traditional person re-ID performs pedestrian retrieval in the full-body person domain. It aims to address the challenge of the large intra-class and small inter-class variation caused by various views, illuminations, poses across disjoint cameras. The mainstream can be grouped into hand-crafted descriptors [35],[42], metric learning methods [3],[16],[18]-[19],[61] and deep learning methods [27]-[29],[44],[60],[62]. In recent years, some research has focused on more practical issues, such as cross-domain [63]-[64], long-term person re-ID [65], and occluded re-ID [7]-[13],[20],[33],[50]. Occluded person re-ID is a challenging practical issue, as occlusion generally occurs in real-world scenarios, but traditional methods usually suffer a dramatic performance drop when dealing with occlusions.

Current works dedicated to occluded person re-ID attempt to seek a matching pattern between local features and global features. Zheng *et al.* [7] proposed a local patch-level matching model named Ambiguity-sensitive Matching Classifier (AMC) and introduced a global part-based matching model. He *et al.* [8] proposed an alignment-free approach named deep spatial feature reconstruction (DSR) to sparsely reconstruct the query images from gallery images. This approach is later improved in [9] and [20] to match different sized feature maps for the occluded re-ID. Zhang *et al.* [43] also introduce a mask-guided de-occlusion (MGD) framework to locate the occlusion and repair the occluded pedestrian, and thus transfer the partial-to-full person matching problem into a full-to-full matching problem. Sun *et al.* [11] introduced a visibility-aware part model (VPM), which learns to perceive the visibility of regions through self-supervision. Zhou *et al*. [13] propose an attention framework to concentrate on the non-occluded region of pedestrians. Luo *et al.* [10] use an affine transform model to transform the holistic image to align with the partial ones. Miao *et al.* [12] propose a pose-guided feature alignment module to match the local patches of query and gallery images based on the human semantic key-points, a benchmark Occlude-DukeMTMC is also proposed, and the method is further improved in [50]. Wang *et al.* [33] also propose a model based on key-points detection and this module can learn high-order relation information for features and topology information for alignment. He *et al.* [5] simultaneously use pose estimation and segmentation to construct the saliency mask for the pedestrian. Besides, [57] uses the Generative Adversarial Networks (GAN) to solve occluded face recognition.

At present, the best performing methods (such as [5],[20], [43], [33]) all use off-the-shelf tools such as instance segmentation and key-points tools. The training processes of these tools need accurate spatial locations or pixel-level annotations, which could be labor-intensive. In this paper, to resist the obstruction in the occluded pedestrian images, instead of using off-the-shelf tools, we seek to make full use of the semantic-level attribute annotations inside person re-ID datasets in a weakly supervised manner, where the annotating process is more labor-saving.

III. Occlusion-sensitive Attribute-guided Body Part Localization

The general architecture of the proposed method is illustrated in Fig. 2. In part A, we describe the CAMs generated from pedestrian attributes. Based on it, the AOPS is introduced in part B.

A. Attribute CAM for Pedestrian Body Part Localization

CAM [4] is a technique to localize the discriminatory image regions even though the network is trained only on image-level labels. But to use CAM, the response categories must have appeared during the training process. And in all re-ID tasks, the ID of the training set and test set are specified as non-overlapping. Therefore, in the person re-ID, it is difficult to use ID tags and CAM to locate the pedestrian parts in the picture.

Lin *et al.* [17] annotated 27 attributes labels for the authoritative person re-ID dataset Market-1501, including gender (male/female), hair (short/long), up sleeve length (short/long), lower body clothing length (short/long), wearing hat (yes/no), age (young/teenager/adult/old), carrying handbag (yes/no), carrying backpack (yes/no), carrying bag (yes/no), 8 colors of upper body clothing and 9 colors of lower-body clothing. These attributes, which cover almost all the characteristics of pedestrians and will appear in both train set and test set, are very suitable for CAM segmentation tasks. However, when there is occlusion in the picture, the attribute CAM is not competent for the task of highlighting the human body parts.

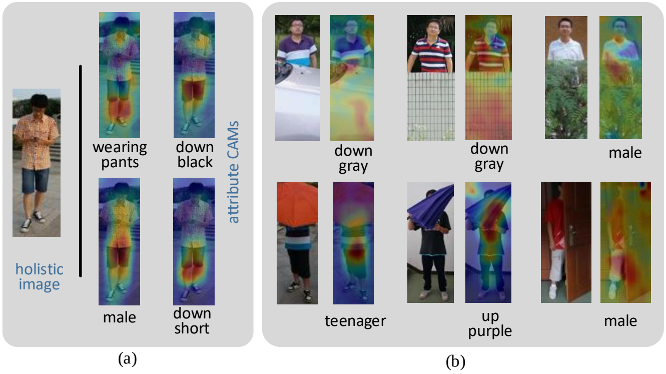

Fig.3. Illustration of attribute CAM. (a) The CAMs of attributes whose prediction probabilities are > 0.9 in a holistic pedestrian image, the region of interest corresponding to each attribute basically conforms to human experience and is concentrated on the human body. (b) Failure cases of attribute CAM when the target person is occluded. When some parts of pedestrians are blocked, under the interference of the obstacles, the CAMs not only appear to be distracting to the background, but also mistake the blocking objects for part of the human body.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

Fig. 3 shows the generated CAMs on pedestrian images using a trained attribute prediction network, the corresponding highlighted area for a full-body image with attribute prediction values greater than 0.9 is drawn. We can see that in Fig. 3 (a), the region of interest corresponding to each attribute basically conforms to human experience and is concentrated on the human body. **However**, in Fig. 3 (b), when some parts of pedestrians are blocked, under the interference of the obstacles, the CAMs not only appear to be distracting to the background, but also mistake the blocking objects for part of the human body. To solve this problem, we propose the interference occlusion erasing strategy to train an AOPS.

B. Attribute-guided Occlusion-sensitive Pedestrian Segmentation

1) Training

Let $\boldsymbol{D}=\{(x_1,l_1,\boldsymbol{a}_1),\dots,(x_N,l_N,\boldsymbol{a}_N)\}$ be the pedestrian training set, where x_i, l_i, and $\boldsymbol{a}_i$ denote the i-th image, its identity label, and its attributes annotations. We can divide $\boldsymbol{D}$ into two parts: $\boldsymbol{D}_L=\{(x_1,l_1),\dots,(x_N,l_N)\}$ and $\boldsymbol{D}_{Attr}=\{(x_1,\boldsymbol{a}_1),\dots,(x_N,\boldsymbol{a}_N)\}$, which denote identity labeled set and attribute labeled set (note that $\boldsymbol{D}_{Attr}$ and $\boldsymbol{D}_L$ share the common pedestrian image $\{x_i\}$). In the training of AOPS, we only use $\boldsymbol{D}_{Attr}$. And for Market-1501 dataset, there are 27 attributes annotated, including 26 attributes with 2 categories (such as gender: male/female, blue up clothing: yes/no, *e.g.*) and 1 attribute with 4 categories (age: young/teenager/adult/ old). For the convenience of using CAM, we formulate the prediction of "age" as 4 binary classification tasks, which are young-yes/no, teenager-yes/no, adult-yes/no, and old-yes/no. Then we have 30 binary classification tasks for attribute learning (*i.e.* $\boldsymbol{a}_i=(a_i^1,a_i^2,\dots,a_i^{30})$, a_i^θ=0 or 1, θ is a the serial number of each of the 30 attributes).

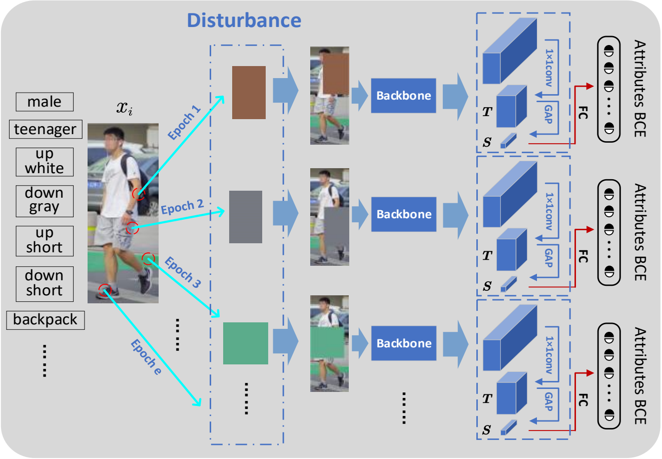

Fig. 4. Training AOPS with interference occlusion erasing strategy. In each epoch, we randomly select a pixel in this image, and randomly select a patch of the picture, and fill the area with this pixel value. The erased image is then input to the backbone network to obtain the feature map.

As is shown in Fig. 4, for one pedestrian image x_i, in each epoch, we randomly select a pixel in this image, and randomly select a patch with an area of 30% to 40% of the picture, and fill the area with this pixel value. This procession is supposed to force the convolution layers to learn the attribute recognition facility under occlusion, thus makes the network to be sensitive to obstructions and focus its attention on the human body. With this disturbance occlusion erased, this image is input to the backbone network to obtain the feature map. Then, a 1×1 convolutional layer is followed to do the dimensionality reduction and we get the feature map $\boldsymbol{T}$, with a size of $c\times h\times w$ (which are the number of channels, height, and width, respectively). Each activation at the spatial location $(\mathcal{X},\mathcal{Y})$ of $\boldsymbol{T}^g$ (a 2-D tensor from $\boldsymbol{T}$, $g\in\{1,\dots,c\}$) is represented by $\mathscr{F}_g(\mathcal{X},\mathcal{Y})$. Then, the result of performing global average pooling (GAP), S^g is:

$$S^g=\sum\nolimits_{\mathcal{X},\mathcal{Y}}\mathscr{F}_g(\mathcal{X},\mathcal{Y}). \tag{1}$$

All S^g are concatenated to $\boldsymbol{S}$, then there are 30 fully connected layers followed by Softmax layers for 30 attributes' learning after $\boldsymbol{S}$. For a certain attribute, there are two categories, assume that the output of its FC layer is $\boldsymbol{z}=[z_1,z_2]$ and the probability of assigning sample x to the attribute class $j\in1,2$ can be written as:

$$p(j|x)=\frac{\exp(z_j)}{\sum_{n=1}^{2}\exp(z_n)}. \tag{2}$$

For brevity, we omit the correlation between j and x. So, the overall binary cross entropy (BCE) loss of attribute classification is formulated as below:

$$\mathcal{L}_{attr}(\boldsymbol{S}_i,\boldsymbol{a}_i)=-\frac{1}{30}\sum\nolimits_{\theta=1}^{30}\sum\nolimits_{j=1}^{2}\log(p(j))q(j). \tag{3}$$

Let y_a be the ground-truth of this attribute label, so that $q(y_a)=1$ and $q(j)=0$ for $j\neq y_a$. θ is a the serial number of each of the 30 attributes. $\boldsymbol{S}_i$ and $\boldsymbol{a}_i$ are the feature descriptor and attribute annotations of x_i, respectively.

2) Body Parts Segmentation

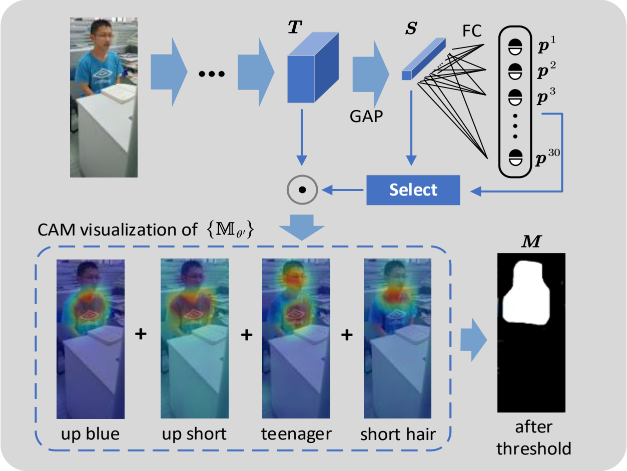

Fig. 5. Locating body part in an occluded person image with AOPS. After training the AOPS with interference occlusion erasing strategy, we can locate the response area on the feature map by observing the predicted values of the attributes, and further locate the position of the human body.

After training, AOPS uses the method illustrated in Fig. 5 to locate the body parts in the occluded pedestrian image. The output of Softmax layer is obtained by Eq. (2), we denote the prediction probabilities of all 30 attributes as $\{\boldsymbol{p}^1,\dots,\boldsymbol{p}^{30}\}$, in which $\boldsymbol{p}^\theta=[p_1^\theta,p_2^\theta]$, $\theta\in\{1,\dots,30\}$. Then, p_2^θ can be regarded as the confidence of the prediction of attribute θ. With

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

a confidence threshold σ empirically set as 0.9, we select a **reliable** attribute set $\{A_{\theta'}\}$ from $\{A_{\theta}\}$, where $\{\theta'\} \in \{\theta\}$ and $p_2^{\theta'} > \sigma$. And for a given $p_2^{\theta'}$, the prediction result of this attribute probability after FC layer is $\sum_g w_g^{\theta'} S^g$, where $w_g^{\theta'}$ is the weight corresponding to the category θ' for the channel g , $g \in \{1, ..., c\}$, S^g is a scalar pooled from $\boldsymbol{T}^g$. Essentially, $w_g^{\theta'}$ indicates the **importance** of $\boldsymbol{T}^g$ for category θ' [4]. Then, several reliable attribute-based heat maps $\{\mathbb{M}_{\theta'}\}$ based on the selected attribute $\{A_{\theta'}\}$ are obtained, where each element is given by:

$$\mathbb{M}_{\theta'}(\mathcal{X},\mathcal{Y}) = \sum_g w_g^{\theta'} \mathscr{F}_g(\mathcal{X},\mathcal{Y}). \tag{4}$$

We perform the visualization of activation maps of attribute CAMs following the method in [4], as is shown in Fig. 5, each map captures a part of the non-occluded human body. Then, β is used to clip the values in maps for discarding background and occluded human part, which is set as 0.7 in our experiments. After that, merging and signing operations are performed, until then a pixel-to-pixel segmentation mask $\boldsymbol{M}$ is available:

$$V_{\theta'}(\mathcal{X},\mathcal{Y}) = sign\left(\mathbb{M}_{\theta'}(\mathcal{X},\mathcal{Y}) - \beta \cdot mean\left(\mathbb{M}_{\theta'}\right)\right), \tag{5}$$

$$\boldsymbol{M}(\mathcal{X},\mathcal{Y}) = sign\left(\sum_{\theta'} V_{\theta'}(\mathcal{X},\mathcal{Y})\right), \tag{6}$$

where the sign operation here sets all non-positive numbers to zero.

IV. Shift Feature Adaption

Since there is currently no natural occluded pedestrian train dataset, we propose the shift feature adaption module to learn the occlusion-sensitive ability using the existing non-occluded person re-ID dataset. As is shown in Fig. 6 (b), in SFA, non-occluded images are first processed with a proposed occlusion batch painting (OBP) strategy, then, a shift attention representation (SAR) takes the segmentation masks and features of artificial occluded images as input to generate the shifted visible features. Finally, these features and corresponding masks are fed into a mask-based drop block (MDB) to refine the representation. Next, we give the details of each component in SFA.

A. Occlusion Batch Painting

As is shown in Fig. 6 (a), to simulate the occlusion, for each batch of training data, we first randomly choose an area as the obstruction region (such as the right half of the image). We fill this area of all **images** in this batch with random color patches, and erase this area of all masks in this batch (*i.e.* fill it with black color). Notable, the OBP should not only be regarded as a data augmentation method for the reason that: In the OBP, on the one hand, the random color painting is supposed to cooperate with the subsequent global pooling (after SAR) to **penalize** SAR when it generates features that contain a large area of obstruction. On the other hand, batch painting and erasing can ensure that all pictures in one batch have the same occlusion area, thereby facilitating the features' triplet metric learning.

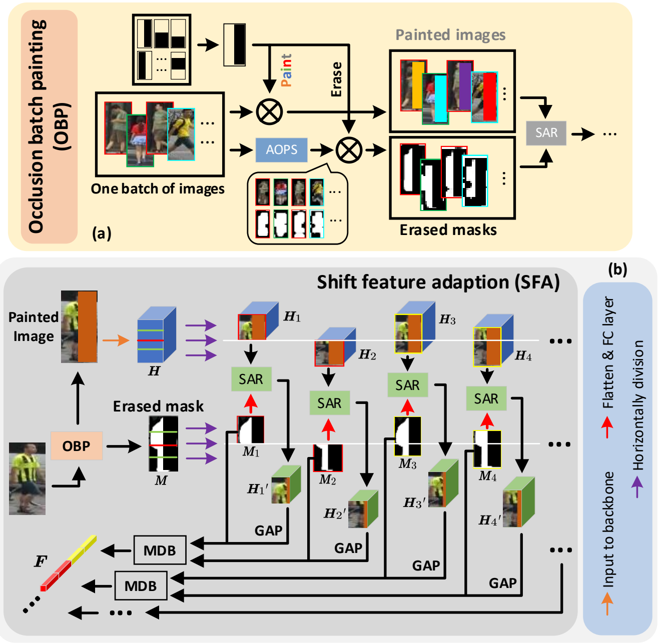

Fig. 6. (a) The illustration of occlusion batch painting (OBP), (b) the illustration of shift feature adaption (SFA). In SFA, non-occluded images are first processed with the occlusion batch painting (OBP) strategy, then, a shift attention representation (SAR) takes the segmentation masks and features of artificial occluded images as input to generate the shifted visible features. Finally, these features and corresponding masks are fed into a mask-based drop block (MDB) to refine the representation.

B. Shift Attention Representation

For a robust representation that does not rely heavily on precise segmentation tools under the occluding scenario, we design the SAR. The SAR only needs to know the approximate position of the obstruction, then it can remove most of the occluder, thereby extracting the discriminative features that contain the visible human body. First, we horizontally divide the color-painted feature $\boldsymbol{H}$ into several parts $\{\boldsymbol{H}_b\}$ to obtain a part-based representation (the details of the division method will be discussed in part E of section VI). $\boldsymbol{M}$ is also divided in the same way into $\{\boldsymbol{M}_b\}$. Then we conduct an affine transformation [26] to $\{\boldsymbol{H}_b\}$, in this transformation, the masks $\{\boldsymbol{M}_b\}$ are processed with a series of operations to serve as the transformation parameters. To make the illustration more concise, we have omitted the following process after each $\boldsymbol{M}_b$ in Fig. 6: $\boldsymbol{M}_b$ is first flattened into a 1-dimensional vector, followed by two fully connected layers, and then mapped into six space transform parameters, formulating a matrix $\mathcal{A}^b$:

$$\mathcal{A}^b = \begin{bmatrix} \phi_1^b & \phi_3^b(=0) & \phi_5^b \\ \phi_4^b(=0) & \phi_2^b & \phi_6^b \end{bmatrix}, \tag{7}$$

where (ϕ_1^b, ϕ_2^b) are scale factors, (ϕ_3^b, ϕ_4^b) are rotation factors, and (ϕ_5^b, ϕ_6^b) control the 2-D spatial position. $\{\boldsymbol{H}_b'\}$ is the set of transformed features. For each pixel $(x_{H'}, y_{H'})$ in $\boldsymbol{H}_b'$, the transforming process from (x_H, y_H) in $\boldsymbol{H}_b$ is given by:

$$\begin{pmatrix} x_{H'}^b \\ y_{H'}^b \end{pmatrix} = \mathcal{A}^b \begin{pmatrix} x_H^b \\ y_H^b \\ 1 \end{pmatrix}. \tag{8}$$

Then, each $\boldsymbol{H}_b'$ is conducted with global average pooling

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

(GAP). In Eq. (7), we set (ϕ_3^b, ϕ_4^b) = (0, 0), which makes the transformation process in Eq. (8) equals a cropping operation to the original feature. The horizontal division to the feature and the cropping transformation of SAR are designed **for the reason** that: the success of the part-based method [44]-[47],[60] indicates that based on the prior knowledge of human body structure, the horizontal splitting of features can be effective for the local information well learned. Moreover, in the non-occluded datasets, we know that even if the image contains a changeable background, the discriminative pedestrian feature can still be learned from the rectangular feature map. Thus, we think that, in the training phase, if we input our model with the pedestrian feature which contains a small part of occlusions (with most of the occlusions are **filtered out**), then as the training proceeds, the model can treat these small amounts of occlusions as part of the background, thereby obtaining a tolerance for small occlusions, and obtain stronger robustness. So, we can see that after SAR in Fig. 6, the network **shifts** its **attention** to the main areas of the human body. Still, these areas ($\boldsymbol{H}_b{'}$) contain small parts of obstructions. Finally, with the following global average pooling to cooperate with OBP, SAR learns a robust representation and avoids heavy dependence on accurate segmentation masks.

C. *Mask-based Drop Block*

Horizontally splitting features can gift SFA more discriminative representations, but in the test set of occluded person re-ID benchmarks, there are often cases where the upper of the lower half of the picture is completely blocked (see (b) of Fig. 3). So, in training, following the real scene, the simulating OBP also contains the situation where the upper or lower half is completely covered. At this time, one of the divided local features could be full of occlusion, and its corresponding mask values will be all-zero. In this case, to prevent SAR from generating meaningless occlusion features to participate in the training, we propose an MDB. The MDB is supposed to suppress the over-occluded features in the identity recognition and metric learning based on the mask values. For a pedestrian image x, its corresponding mask is split into $\{\boldsymbol{M}_b\}$, the feature map $\boldsymbol{H}$ is divided into $\{\boldsymbol{H}_b\}$ and then transformed to $\{\boldsymbol{H}_b{'}\}$, which are pooled to $\{(\boldsymbol{H}_b{'})^p\}$. The procession of MDB is denoted as:

$$(\boldsymbol{H}_b{'})^p \leftarrow \frac{nonezeros(\boldsymbol{M_b})}{nonezeros(\boldsymbol{M_b}) + zeros(\boldsymbol{M_b})} (\boldsymbol{H}_b{'})^p . \quad (9)$$

The feature vectors performed with MDP are concatenated together as the final descriptor of the pedestrian image.

V. Loss Function and VRM

A. *Training Loss*

At the end of the SFA module, the descriptor $\boldsymbol{F}$ is used to perform the metric learning and identity classification learning. For metric learning [3], in one batch, we randomly select $\mathcal{P}$ persons and pick $\mathcal{K}$ images of each person, *i.e.* totally $\mathcal{PK}$ images. Our goal is to make the distance between features of the same ID smaller than the distance between features of different IDs. Given a training image x_i^l whose ID is l, its feature descriptor is $\boldsymbol{F}_i^l$, then, descriptors $\boldsymbol{F}_f^l$ for all $f \neq i$ are regarded as positive examples $\{\boldsymbol{F}_i^{(pos)}\}$, and for all $\boldsymbol{F}_j^e$ that $e \neq l$ are negative examples $\{\boldsymbol{F}_i^{(neg)}\}$. For each $\boldsymbol{F}_i^l$ in this batch, we find its hardest positive and negative example, the hard example mining triplet loss [3] is given by:

$$\mathcal{L}_{HEM} = \overbrace{\sum_{i=1}^{\mathcal{PK}}}^{all\ anchors} \Bigg[\tau + \overbrace{\max_{i=1\ldots\mathcal{PK}} Eu\Big(\boldsymbol{F}_i, \boldsymbol{F}_i^{(pos)}\Big)}^{hardest\ positive} - \underbrace{\min_{i=1\ldots\mathcal{PK}} Eu\Big(\boldsymbol{F}_i, \boldsymbol{F}_i^{(neg)}\Big)}_{hardest\ negative} \Bigg]_+ , \quad (10)$$

where k means the kth image from $\mathcal{K}$ images of one person, $Eu(\bullet)$ means Euclidean Distance calculation, and τ is the margin enforced between positive and negative examples.

Simultaneously, $\boldsymbol{F}_i^l$ is also imposed with an identity classification loss. Assume that the output of FC in ID classifiers is $\boldsymbol{v} = [v_1, \ldots, v_I]$. The predicted probability of each ID label n is calculated as:

$$p(n|x) = \frac{\exp(v_n)}{\sum_{\varphi=1}^{I} \exp(v_\varphi)} . \quad (11)$$

The cross-entropy loss of ID classification is formulated as:

$$\mathcal{L}_{ID}\Big(\boldsymbol{F}_i^l, l\Big) = -\sum_{n=1}^{I} \log(p(n)) q(n) . \quad (12)$$

Let y_l be the ground-truth ID label, so that $q(y_l) = 1$ and $q(n) = 0$ for all $n \neq y_l$. In this case, minimizing the cross entropy is equivalent to maximizing the possibility of being classified to the ground-truth category. Thus, the final loss of SFA is:

$$\mathcal{L}_{SFA} = \mathcal{L}_{HEM} + \gamma \cdot \mathcal{L}_{ID} . \quad (13)$$

Parameter γ balances the contributions of these two losses.

B. *Visible Region Matching*

Algorithm 1 Computing distance in the retrieval process with visible region matching (VRM).

```
Input: One query (occluded) image $I_q$ and all gallery
    (holistic) images $\{I_{g1}, I_{g2}, \ldots\}$.
Output: Euclidean Distances of one query image to each
    gallery image.
1. Use AOPS to predict attributes of $I_q$.
2. Solve Eq. (4), (5), (6) to obtain segmentation mask $\boldsymbol{M}_q$.
3. Utilize $\boldsymbol{M}_q$ as the segmentation masks of all gallery
    images $\boldsymbol{M}_{g1} \leftarrow \boldsymbol{M}_q$, $\boldsymbol{M}_{g2} \leftarrow \boldsymbol{M}_q$, ...
4. Input masks $\boldsymbol{M}_q$, $\boldsymbol{M}_{g1}$, $\boldsymbol{M}_{g2} \ldots$ together with feature
    maps $\boldsymbol{H}_q$, $\boldsymbol{H}_{g1}$, $\boldsymbol{H}_{g2} \ldots$ to SFA, and get the descriptors
    $\boldsymbol{F}_q$, $\boldsymbol{F}_{g1}$, $\boldsymbol{F}_{g2}$, $\ldots$
5. Compute the Euclidean Distance of $\boldsymbol{F}_q$ to each of $\boldsymbol{F}_{g1}$,
    $\boldsymbol{F}_{g2} \ldots$ : $Eu\langle \boldsymbol{F}_q, \boldsymbol{F}_{g1} \rangle, Eu\langle \boldsymbol{F}_q, \boldsymbol{F}_{g2} \rangle$, $\ldots$
```

AOPS locates the pedestrian position in the occluded image. With the localization mask from AOPS, SFA trains a network to extract corresponding discriminative features. Based on these modules, we design the visible region matching (VRM) method to overcome the interference caused by useless parts of the full-

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

body picture (mentioned in Fig. 1) during the retrieval phase. The main idea of VRM is to replace the segmentation mask of all the searched objects (taken from gallery set and usually unobstructed) with the mask of the current target person image (taken from query set and there is occlusion) in each retrieval process. The VRM is illustrated in Algorithm 1, and it mainly describes the calculation method of the Euclidean distance between the query and gallery in our framework. In this way, the VRM guides the SFA to generate effective representation from the area of gallery images that is the same as the position of the human body locates in the query image.

VI. Experiments

In this section, we first introduce the datasets used and the model implementation details in part A and part B. Then, the comparisons with the-state-of-the-art occluded person re-ID methods are described in section C. To verify the effectiveness of the components in our method, the ablation studies are presented in part D-F in the order of AOPS, SFA, and VRM. The experiments on occluded vehicle re-ID are shown in part G. Finally, the visualization of retrieval results is discussed in part H.

A. Datasets and Evaluation Protocol

We evaluate our method on four occluded/partial person re-ID datasets: Partial-REID [7], Partial-iLIDS [21], and Occluded-REID [13]. To validate the generalization ability of our framework, we also conduct experiments on two holistic datasets Market-1501 [22], DukeMTMC-reID [23], and a vehicle re-ID dataset VeRi-776 [30][49]. Partial-REID and Partial-iLIDS are two commonly-used datasets for partial re-ID, but the former also has the occluded version for occluded re-ID. Partial-REID contains 600 images and 60 identities, each one of which has 5 occluded images and 5 holistic images. Partial-iLIDS has 238 images from 119 identities. Occluded-REID is an occluded person dataset that consists of 2000 images of 200 occluded persons. Each identity has 5 occluded images and 5 holistic images. Market-1501 and DukeMTMC-reID both contain few occluded person images and are widely treated as a holistic re-ID dataset. Market-1501 consists of 32,688 images of 1501 subjects observed from 6 camera viewpoints. Its training set, gallery set, and query set respectively contains 12,936, 19,732, and 3,368 images. DukeMTMC-reID dataset contains 1,404 identities, 16,522 training images, 2,228 queries, and 17,661 gallery images. Occluded-DukeMTMC is selected from DukeMTMC-reID by leaving occluded images and filter out some overlap images. It contains 15,618 training images, 17,661 gallery images, and 2,210 occluded query images. VeRi-776 contains 49,357 images of 776 vehicles from 20 cameras. This dataset is collected in a natural traffic environment. The vehicles are labeled with bounding boxes over the whole vehicle body, type, color, and cross camera vehicle correlation.

It should be noted that as mentioned in section I, we are focusing on re-identification of occluded person re-ID. However, in the partial person re-ID dataset Partial-iLIDS, all pedestrian obstructions are suitcases, and all the examples in the query set are pictures of only a part of the pedestrian's body obtained by manually cutting the suitcases. In order to simulate the occluded situation, for the query set of this Partial-iLIDS, we add a color block similar as shown in Fig. 7 below each image, the color block area accounts for 1/3 of the whole picture, and we name the modified dataset as **Partial_iLIDS_O**.

We use cumulative matching characteristic (CMC, also known as top-K accuracy) curves and mean average precision (mAP) to evaluate different models in our experiments. We follow the evaluation protocol in [8], the occluded images and holistic images in occluded person re-ID datasets are respectively regarded as query and gallery. For vehicle re-ID, since there is no special occluded dataset, we follow the setting of Partial-REID, manually select 60 vehicles from the VeRi-776 test set, each vehicle has 5 query images and 5 gallery images, and then block the query set with random colors to simulate the occlusion. All experiments are performed in the single query setting.

B. Implementation Details

We use ResNet50 [24] as our backbone, then we remove the last global average pooling layer and fully connected layer. The initialized model is pre-trained by ImageNet [25]. Input images are resized to 384×128 and augmented by flipping the picture horizontally with a probability of 50%. In one epoch, each picture is augmented once. We set the batch size to 64 (*i.e.* PK = 64, including 16 identities, 4 images per identity). In AOPS, the feature map after the backbone is performed 1×1 convolution to reduce the dimensionality from 2048 to 512. Stochastic gradient descent is applied with a momentum of 0.9. The learning rate is gradually increased from 1e-5 to 1e-3 in total of 80 training epochs. The AOPSs used on occluded datasets are all trained on Market-1501. In SFA, following [33], we train the module on Market-1501 and use color jitter augmentation to avoid domain variance when test on occluded datasets. We use Adam optimizer [31] with the base learning rate initialized to 1.2e-3 with a linear warm-up [32] in the first 20 epochs, then decayed to its 0.1 in 40 and 70 epochs, a total of 180 training epochs are required. In the validation on the vehicle dataset, the Adam optimizer is set with the base learning rate initialized to 1e-3 with a linear warm-up in the first 20 epochs, then decayed to its 0.6 every 15 epochs between 20 and 100 epochs. The GPU we utilize is NVIDIA RTX 2080Ti, we use Pytorch 1.0 to establish our whole framework. Based on the ablation studies in part D-E, the super parameters of our method in part C are set τ =0.3 (in Eq. (10)), γ =2 (in Eq. (13)). And the feature map division method in SFA is set Mode O_2+N_2 introduced in part E, 1). Additionally, to implement the variable-controlled approach, in each ablation study section corresponding to certain super-parameters, other super-parameters set the values which lead to the best performance.

C. Comparison with the state-of-the-art

In this part, 11 existing occluded person re-ID methods that have been introduced in section II are used for comparison, including AMC+SWM (ICCV15') [7], DSR [8] (CVPR18'), SFR [9], STNReID [10], VPM (CVPR19') [11], PGFA (ICCV19') [12], PGFA+ [50] (TNNLS21', the journal version of PGFA), AFPB (ICME18') [13], FPR (ICCV19') [20], MGD (IScIDE19') [43], HONet (CVPR20') [33], GASM [5]

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I
The Comparison of Our Method with the State-of-the-art Methods on Partial-REID Dataset, Rank-1 Accuracy (%), Rank-3 Accuracy (%), Rank-5 Accuracy (%), and mAP (%) Are Shown.

Methods	Partial-REID			
	R1	R3	R5	mAP
AMC+SWM [7]	37.3	46.0	-	-
DSR [8]	50.7	70.3	-	-
SFR [9]	56.9	78.5	-	-
STNReID [10]	66.7	80.3	86.0	-
VPM [11]	67.7	81.9	-	-
PGFA [12]	68.0	80.0	82.0	56.2
PGFA+ [50]	72.5	83.0	-	-
AFPB [13]	78.5	-	94.9	-
FPR [20]	81.0	-	-	76.6
MGD [43]	84.3	-	94.0	-
HONet [33]	85.3	91.0	-	-
Baseline	64.9	75.3	81.7	56.5
AOPS +SFA	83.2	89.6	93.2	76.0
AOPS +SFA+VRM	**86.8**	**93.5**	**95.5**	**78.8**

TABLE II
The Comparison of Our Method with the State-of-the-art Methods on Partial-iLIDS and Partial-iLIDS_O Datasets, Rank-1 Accuracy (%), Rank-3 Accuracy (%), Rank-5 Accuracy (%), and mAP (%) Are Shown.

Methods	Partial-iLIDS			
	R1	R3	R5	mAP
AWC+SWM [7]	21.0	32.8	-	-
DSR [8]	58.8	67.2	-	-
SFR [9]	63.9	74.8	-	-
STNReID [10]	54.6	71.3	79.2	
VPM [11]	67.2	76.5	-	-
PGFA [12]	69.1	80.9	-	-
FPR [20]	68.1	-	-	61.8
PGFA+ [50]	70.6	81.3	-	-
HONet [33]	**72.6**	**86.4**	-	-
Baseline	56.3	65.5	69.1	49.0
AOPS +SFA	71.4	81.9	**83.0**	**72.5**
Method	Partial-iLIDS_O			
Baseline	67.2	71.9	76.2	62.5
AOPS +SFA	79.0	85.3	87.7	80.8
AOPS +SFA+VRM	**81.7**	**88.3**	**90.9**	85.9

(ECCV20'). The person re-ID accuracy comparison with the state-of-the-art methods on Partial-REID, Partial-iLIDS, Occluded REID, and Occluded-DukeMTMC are shown in Table I-IV. As we can see, the three best performing competitors are FPR [20], PGFA+ [50], and HONet [33]. As is mentioned in section II, to locate the human body part in the picture, they borrow the read-made human segmentation tool or key-point detection tool. But the segmentation method AOPS in our method is a weakly supervised trained model, which makes it more valuable for us to reach the same performance level as these methods.

As shown in Table I and Table III, on Partial-REID and Occluded REID, our AOPS+SFA+VRM outperforms all the existing state-of-the-art in both rank-1 and mAP. On Partial-REID, our method AOPS+SFA+VRM has advantages of +(1.5%/2.5%) in rank-1/rank-3 over the strongest competitor HONet. And on Occluded REID, this advantage of AOPS+SFA +VRM over HONet is +(2.2%/1.6%). On Partial-iLIDS, since there are no occluded pedestrian pictures, VRM is not suitable for this dataset. But in Table II we can see that when only using AOPS+SFA on Partial-iLIDS, our method still outperforms all methods except for HONet. On the modified Partial-iLIDS_O, AOPS+SFA gains improvement of +(7.6%/8.3%) in rank-1/mAP over on Partial-iLIDS. This is because our method is designed specifically for the occluded person re-ID. Considering that it is labor-intensive to manually crop the occlusions in the actual application of re-ID, we value the results on the occluded dataset Partial-iLIDS_O rather than the partial dataset Partial-iLIDS. Then, with VRM, our rank-1 on Partial-iLIDS_O has been significantly improved to 81.7%, which is 9.1% higher than the best accuracy (HONet) on Partial-iLIDS.

TABLE III
The Comparison of Our Method with the State-of-the-art Methods on Occluded REID, Rank-1 Accuracy (%), Rank-3 Accuracy (%), Rank-5 Accuracy (%), and mAP (%) Are Shown.

Methods	Occluded REID			
	R1	R3	R5	mAP
AWC+SWM [7]	31.1	-	-	27.3
PCB [44]	41.3	-	-	38.9
DSR [8]	72.8	-	-	62.8
GASM [5]	74.5	-	-	65.6
FPR [20]	78.3	-	-	68.0
HONet [33]	80.3	-	-	70.2
Baseline	58.3	69.6	75.5	49.4
AOPS+SFA	80.3	86.1	88.3	70.1
AOPS +SFA+VRM	**82.5**	**89.7**	**92.2**	**71.8**

TABLE IV
The Comparison of Our Method with the State-of-the-art Methods on Occluded-DukeMTMC, Rank-1 Accuracy (%), Rank-5 Accuracy (%), Rank-10 Accuracy (%), and mAP (%) Are Shown.

Methods	Occluded-DukeMTMC			
	R1	R5	R10	mAP
PCB [44]	42.6	57.1	62.9	33.7
DSR [8]	40.8	58.2	65.2	30.4
SFR [9]	42.3	60.3	67.3	32.0
PGFA [12]	51.4	68.6	74.9	37.3
HONet [33]	55.1	-	-	**43.8**
PGFA+ [50]	**56.3**	72.4	78.0	43.5
Baseline	39.9	55.8	60.0	31.7
AOPS+SFA	54.0	70.3	77.2	42.6
AOPS +SFA+VRM	55.4	**72.4**	**78.9**	**43.8**

In order to show the effectiveness of our method more intuitively, we also construct a **baseline** for comparison, we remove the proposed parts from our framework, leaving only Resnet50 as the feature extraction network, and then use ID loss and hard example mining triplet as the training loss. We can see that the improvement of our method is obvious on the three datasets, for AOPS+SFA+VRM over baseline, the improvements are +(11.9%/22.3%), +(14.5%/23.4%), and +(24.2%/22.4%) on Partial-REID, Partial-iLIDS_O, and Occluded REID in rank-1/mAP, respectively.

Occluded-DukeMTMC is proposed by PGFA+ [50], different from the other three occluded datasets, the gallery of this dataset contains 10% occluded images, and [50] also designed an effective algorithm for this situation and achieved encouraging performance. As shown in Table IV, our AOPS+ SFA has achieved good performance by filtering out occlusions and learning robust part-based feature representation. Based on AOPS+SFA, VRM can still gain a performance improvement of +1.4%/1.2% in rank-1/mAP by solving the interference problem illustrated in Fig. 1. However, because a small number of pictures in the gallery are occluded, and the occlusion parts are not always the same as the query, the improvement of VRM on Occluded-DukeMTMC is not as obvious as that of the other three datasets. But in the end, our performance has also reached

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

the state-of-the-art, which indicates a good universality of our method on this challenging benchmark.

TABLE V
THE COMPARISON OF OUR METHOD WITH THE OCCLUDED PERSON RE-ID STATE-OF-THE-ART METHODS ON MARKET-1501 AND DUKEMTMC-REID, RANK-1 ACCURACY (%) AND MAP (%) ARE SHOWN.

Methods	Market-1501		DukeMTMC-reID	
	R1	mAP	R1	mAP
DSR [8]	91.3	75.6	82.5	68.7
SFR [9]	93.1	81.0	84.9	71.3
VPM [11]	93.0	80.8	83.6	72.6
PGFA [12]	91.2	76.8	82.6	65.5
FPR [20]	**95.4**	**86.6**	**88.6**	**78.4**
HONet [33]	94.2	84.9	86.9	75.6
GASM [5]	95.3	84.7	88.3	74.4
Baseline	93.4	84.1	86.2	74.1
AOPS+SFA	94.6	85.2	87.4	76.3
AOPS+SFA+VRM	94.6	85.3	87.5	76.3

Since the occluded re-ID methods are tailor-made for the scenario where pedestrians are occluded, the latest occluded pedestrian re-ID methods, in addition to testing on the occluded dataset, usually also observe the universality on the holistic pedestrian datasets. We follow them and test our method on the two most popular holistic datasets Market-1501 and Duke-MTMC-reID. As shown in Table V, FPR [20] achieves much higher performance than other methods, this may be because the picture descriptor of FPR is designed as four complex feature maps instead of the concatenated feature vectors used by most methods (each feature map can be equivalent to multiple feature vectors). Therefore, while consuming more storage and computing resources, FPR contains richer information and performs well on large holistic datasets. And our mAP ranks second among all occluded re-ID methods, which indicates a good generality of our method on holistic datasets. We can also see that, since there is few obstruction on the image of holistic dataset, the gains of VRM are very weak. And the gains of AOPS+SFA over baseline are +(1.2%/1.1%) and +(1.2%/1.2%) in rank-1/mAP on two holistic datasets, this improvement indicates that even if the occlusion is scarce in the dataset, the part-based representation of AOPS+SFA can still be helpful for enhancing the local features' learning on the basic backbone.

D. Ablation Study on AOPS

1) Comparison with Standard Attribute Network

In our framework, AOPS is responsible for locating the pedestrian body part, and its role is like a human parsing or segmentation tool. But it is inappropriate to use intersection over union (IoU) to evaluate AOPS like a segmentation method. Because AOPS is weakly supervised trained, it cannot provide an extremely accurate segmentation mask. Besides, in our system, we don't need it to provide a mask that can accurately cut the pedestrian's body part: In SFA, we fixed the rotation factors of the spatial transformation, the extracted local features are a rectangular region derived from the overall features. The human body is non-rigid, so that local features must contain a small amount of **occlusion and background information**. In the training stage, we let the model recognize pedestrian images that contain some background information, which enables the network to construct more robust discrimination. Then, for the AOPS, it only needs to provide the SFA with guidance information on where in the image the pedestrian's body parts **roughly** locate.

Therefore, we designed the following experiment to verify the effectiveness of AOPS: we first train a standard attribute recognition network (named **AttrNet**) on Market-1501 with the attribute annotations, and then train the AOPS with interference occlusion erasing strategy shown in Fig. 4. Next, we fill the left half or right half random color blocks (such as yellow, green, blue, *etc.*) of the pictures in the test set. The reason why we choose left and right instead of up and down is that the left and right half of the fill has less effect on the original pedestrian's attributes. For example, if the lower body of a pedestrian is blocked, the color of the lower body is not visible, and this lower body color attribute annotated will disappear. But if we block the left half, this will not happen. We call the test set filled with color patches as Halfpatint-Testset, and then we use AttrNet and AOPS to test the attribute prediction accuracy on the original test set (Ori-Testset) and Halfpatint-Testset, respectively. Fig. 7 shows some CAM heat maps examples of these two methods.

It can be seen that when AttrNet handles the situation where the color block occludes the pedestrian's body, the attribute prediction will be disturbed, and some attention will be dispersed to the color block. The color of the painted block will not only cause AttrNet to predict the wrong color attribute (as shown in Fig. 7 (a) and (b)), it will also interfere with the network's judgment of other non-color attributes (Figure 7 (c)). However, AOPS will focus on the human body part to avoid interference of color blocks. And compared with AttrNet, AOPS can focus on the part of the picture that is more in line with the human prior knowledge of each attribute.

We then follow the attribute accuracy evaluation method of [17], using mean accuracy to observe the attribute prediction

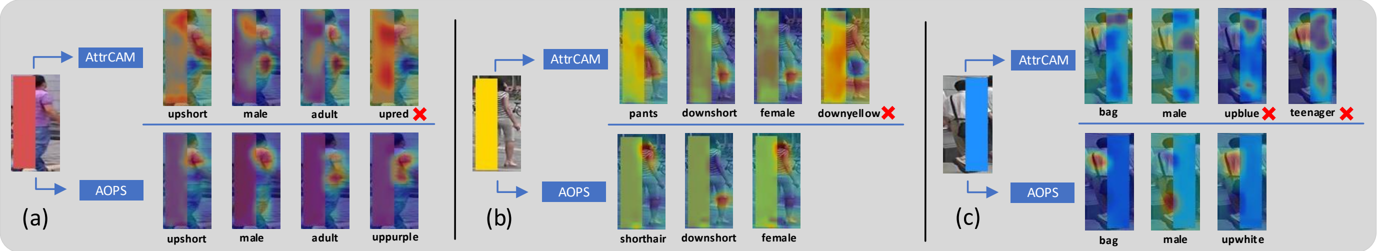

Fig.7. CAM cases when AttrNet and AOPS are tested on Halfpatint-Testset (transformed from Market-1501). For one image, only the CAMs of attributes who have prediction probabilities greater than 0.9 are shown. It can be seen that when AttrNet handles the situation where the color block occludes the pedestrian's body, the attribute prediction will be disturbed, and some attention will be even dispersed to the color block. And compared with AttrNet, AOPS can focus on the part of the picture that is more in line with the human prior knowledge of each attribute.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

performance of the two methods (the threshold for judging whether the prediction probability is correct is set to 0.5 in this evaluation). Table VI shows the mean attribute prediction accuracies of 30 attributes of the two methods on different test sets. We can see that when tested on the original test set (Ori-Testset), AOPS has a 0.5% performance advantage over AttrNet. This is because even tested on images without occlusion, AOPS can focus more on the correct position than AttrNet does (see the CAMs of "upshort" attribute of AttrNet and AOPS in Fig. 7 (a)). When the two methods were tested on the dataset filled with colored blocks (Halfpatint-Testset), the accuracy of AttrNet decreases from 91.6% to 82.6% (a 9% drop), while the performance of AOPS only decrease by 1.2%.

TABLE VI
THE ATTRIBUTE PREDICTION ACCURACIES OF ATTRNET AND PROPOSED AOPS ON ORIGINAL TESTSET AND COLOR HALF-PAINTED TESTSET (BOTH FROM MARKET-1501), MEAN ACCURACY ARE SHOWN.

Methods	Attribute Mean Accuracy	
	Ori-Testset	Halfpatint-Testset
AttrNet	91.6%	82.6%
AOPS	**92.1%**	**90.9%**

2) *Comparison with Human Parsing Network*

We also conduct another interesting set of experiments. We use an existing human parsing model CE2P [48] instead of AOPS, and AttrNet instead of AOPS, combined with SFA to complete the task of occluded person re-ID. The re-ID experimental results on the three occluded person re-ID datasets are shown in Table VII. It can be seen that AttrNet+SFA has the worst performance, which is easy to foresee. But the experimental results of AOPS+SFA+VRM and CE2P+SFA+ VRM are quite close. It should be known that, CE2P is a strongly supervised segmentation model trained with pixel-level annotations. It segments pedestrian body parts from the background much more accurately than AOPS. But when imposed to our framework together with SFA, it helps achieve similar re-ID performance as AOPS does. This is because, in SFA, feature extraction does not need accurate segmentation maps (as mentioned in paragraph 1 in section D). The model has seen many pedestrian pictures with a small amount of background or even obstructions during the training phase. In the test phase, even if the local feature extracted contains a small amount of interference information, the network can still identity the pedestrian ID correctly.

TABLE VII
THE COMPARISON OF REPLACING AOPS WITH ATTRNET AND CE2P IN OUR OCCLUDED PERSON RE-IDENTIFICATION FRAMEWORK, RESULTS ON PARTIAL-REID, OCCLUDED-DUKEMTMC AND OCCLUDED REID ARE REPORTED, RANK-1 ACCURACY (%) AND MAP (%) ARE SHOWN.

Methods	Partial-REID		Occluded-DukeMTMC		Occluded REID	
	R1	mAP	R1	mAP	R1	mAP
AttrNet+SFA+VRM	66.5	57.5	33.6	23.3	62.8	49.3
AOPS+SFA+VRM	86.8	**78.8**	**55.4**	43.8	**82.5**	71.8
CE2P+SFA+VRM	**86.9**	78.7	**55.4**	**43.9**	82.4	**71.9**

3) *Parameter Analysis*

In the part B, 2) of section III, we select the **reliable** attributes set $\{A_{\theta'}\}$ from $\{A_{\theta}\}$ with a confidence threshold σ, where the prediction value of these attributes $p_2^{\theta'} > \sigma$. Parameter σ determines the strictness of the selection of attributes when we use attributes to locate the human body. Since the attribute prediction is a binary classification task, $\sigma \in [0.5, 1)$. Table VIII shows the performance changes of our method on the three datasets as σ changes. It can be seen that when the σ is changed from 0.5 to 0.7, the performance is gradually improved. The results of 0.8 and 0.9 for are similar. This may be because some backgrounds or occlusions similar to the human body will be recognized as attributes by the AOPS with lower confidence. When the value of σ is set small, it will affect the accuracy of human parsing. When the σ value is greater than a certain value, the response area is more reliable for attributes predicted with high confidence. The small fluctuation of σ will not bring about big performance fluctuations.

TABLE VIII
THE NUMERICAL RESULTS UNDER DIFFERENT VALUES OF σ IN AOPS, RESULTS ON PARTIAL-REID, OCCLUDED-DUKEMTMC AND OCCLUDED REID ARE REPORTED, RANK-1 ACCURACY (%) AND MAP (%) ARE SHOWN.

Value of σ	Partial-REID		Occluded-DukeMTMC		Occluded REID	
	R1	mAP	R1	mAP	R1	mAP
0.5	83.7	72.6	50.9	38.7	77.9	67.2
0.6	85.2	75.5	53.1	41.0	79.1	69.0
0.7	86.0	77.6	54.2	42.1	80.9	70.6
0.8	86.7	78.8	55.4	43.7	82.5	71.8
0.9	86.8	78.8	55.4	43.8	82.5	71.8

The method of selecting attributes based on thresholds in AOPS is somewhat like self-paced learning in unsupervised methods [51][52][54]. As the training progresses, the recognition ability of the network will gradually improve. But the filtering is performed after the training is completed, and the samples in our dataset are fully-labelled. For the training process of AOPS, Fig. 8 shows the changes in loss and accuracy as the epoch changes. The attribute prediction accuracy is tested on Halfpatint-Testset introduced in Table VI. We can see that as training progresses, the training loss continues to decrease, and performance gradually rises with small fluctuations. Finally, it stabilizes at 90.9% at the 80th epoch. Under our experimental conditions, complete training only needs about 4 hours to cycle the Market-1501 dataset for 80 epochs.

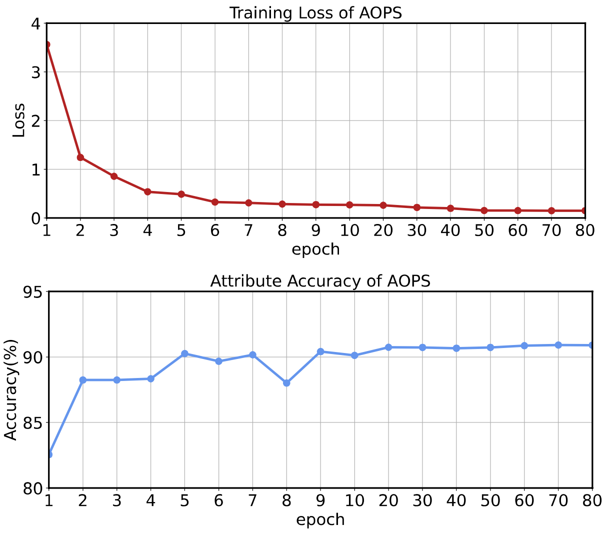

Fig.8. The training loss (on train set) and attribute prediction accuracy (on test set) curves of AOPS. As training progresses, loss continues to decrease, and performance stabilizes at 90.9% at the 80th epoch.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

E. Ablation Study on SFA

In this section, the feature adaption module is analyzed from aspects of feature division, SAR parameter setting, loss function parameter setting, and effectiveness validation.

1) Local Feature Division

In this part, we study the division method of features before SAR. To obtain a richer granularity of feature representation, we formulate two different types of local features: overlapping features and non-overlapping features. For overlapping features that cover a larger area in the image, cutting the feature map into two parts in two different ways: the upper and lower 2/3 of image (Mode O_1), the upper and lower 3/4 of the image (Mode O_2). Then for non-overlapping features which have smaller receptive field, we follow the division methods in PCB [44][60], horizontally divide the feature map into 2 stripes (Mode N_1), 3 stripes (Mode N_2), and 6 stripes (Mode N_3) on average. These overlapping and non-overlapping features serve as local representations. Inspired by MGN [28], we also add a global representation Mode G. Since we do not use a separate ID loss for each granularity like MGN, we only add one global vector. The results of AOPS+SFA+VRM are reported in Table IX.

TABLE IX
THE NUMERICAL RESULTS OF DIFFERENT FEATURE DIVISION MODE IN SFA, RESULTS ON PARTIAL-REID, OCCLUDED-DUKEMTMC AND OCCLUDED REID ARE REPORTED, RANK-1 ACCURACY (%) AND MAP (%) ARE SHOWN.

Methods	Partial-REID		Occluded-Duke		Occluded REID	
	R1	mAP	R1	mAP	R1	mAP
Mode G	71.6	62.7	43.3	34.5	62.4	52.5
Mode O_1	81.1	71.2	51.5	39.9	77.9	66.8
Mode O_2	82.8	75.0	52.8	41.0	78.7	68.1
Mode N_1	81.6	71.3	51.0	39.6	77.6	63.8
Mode N_2	82.7	74.3	52.7	41.1	78.8	66.8
Mode N_3	82.2	72.5	51.9	40.3	78.5	66.0
Mode O_1+N_1	82.6	76.0	52.5	40.8	79.2	68.6
Mode O_1+N_2	85.3	77.5	54.4	42.7	81.5	70.6
Mode O_1+N_3	83.7	76.8	53.2	41.4	80.0	69.4
Mode O_2+N_1	84.1	76.4	53.8	42.0	79.6	68.8
Mode O_2+N_2	**86.6**	**78.7**	**55.2**	**43.7**	**82.3**	**71.7**
Mode O_2+N_3	85.5	77.2	54.6	43.0	80.7	70.1
Mode O_2+N_2+G	**86.8**	**78.8**	**55.4**	**43.8**	**82.5**	**71.8**

It can be seen that when only overlapping features are used, the upper and lower 3/4 block method (Mode O_2) achieves the best performance, and among the non-overlapping block methods, the three-part method (Mode N_2) has the best performance. We combine these two types of features together, then the best method is Mode O_2+N_2, the dual-granularity local representation brings a performance gain of +(3.8%/3.7%) in rank-1/mAP on Partial-REID on the basis of Mode O_2. On Occluded-DukeMTMC and Occluded REID the gains are +(2.4%/2.7%) and +(3.6%/3.6%). Then, we add the global representation, which also brings gains of +(0.2%/0.1%), +(0.2%/0.1%), and +(0.2%/0.1%).

It should be noted that the results in Table IX are only based on our artificial division method. In [53], the authors explored the use of neural architecture search (NAS) technology to find a better part model for person re-ID. Therefore, if the experimental conditions are sufficient, it is possible to search for a better-performing part-based structure for occluded re-ID with NAS.

2) SAR Parameter Analysis

In SFA, SAR extracts features based on the pedestrian mask provided by AOPS. In SAR, there are six parameters that control the affine transformation, and they can be divided into two types: the parameters (ϕ_1^b, ϕ_2^b), (ϕ_5^b, ϕ_6^b) that implement the translation and cropping function, and the parameters (ϕ_3^p, ϕ_4^p) that implement the rotation function. Note that when only the rotation function is used, the feature cannot be cut, so that the obstruction cannot be avoided. Therefore, we ignore the situation using only (ϕ_3^p, ϕ_4^p), and conduct two sets of experiments. In setting ***a***, only translation and cutting are valid, *i.e.*, the parameters to be learned are (ϕ_1^b, ϕ_2^b) and (ϕ_5^b, ϕ_6^b). And in setting ***b***, all six parameters are to be learned. The experimental results are shown in Table X. The two sets of experiments are denoted as AOPS+SFA+VRM with six parameters (*w.* 6 *para.*) and with four parameters (*w.* 4 *para.*).

TABLE X
THE NUMERICAL RESULTS OF DIFFERENT PARAMETER SETTING IN SFA, RESULTS ON PARTIAL-REID, OCCLUDED-DUKEMTMC AND OCCLUDED REID ARE REPORTED, RANK-1 ACCURACY (%) AND MAP (%) ARE SHOWN.

Methods	Partial-REID		Occluded-Duke		Occluded REID	
	R1	mAP	R1	mAP	R1	mAP
AOPS+SFA+VRM *w.* 6 *para.*	83.9	75.4	52.1	41.4	79.3	68.4
AOPS+SFA +VRM *w.* 4 *para.*	**86.8**	**78.8**	**55.4**	**43.8**	**82.5**	**71.8**

We can see that when learning only translation and cutting parameters, it can achieve a better performance (AOPS+SFA +VRM *w.* 4 *para.*). This may be because rotating pedestrians actually does not fit the way pedestrian recognition in our experience, and keeping the vertical direction of the pedestrian makes the local feature learning more stable. On the other hand, in the case of only the cutting and translation are valid, as shown in Fig. 6, the extracted feature is a small rectangular area from the original feature, which contains a few occlusion and background information. This helps the model improve its robustness against background interference. In the test phase, even if the mask cannot accurately divide the pedestrians from the occlusion, the well-learned model can still identity pedestrian features that contain little obstructions and background, thereby achieving higher re-ID accuracy.

3) Parameter Analysis of Loss Function

In Eq. (10), parameter τ determines the margin that is enforced between positive and negative pairs, and in Eq. (13), γ balances the contributions of triplet loss and ID cross entropy loss. We set the value of τ from 0.1 to 1.0 at the stride of 0.1, and the value of γ from 0.5 to 3.5 at the stride of 0.5. Fig. 9 shows the rank-1 and mAP curves that vary with these two parameters. We can see that the performance changes are similar on the three datasets. For the margin in triplet loss, when τ is 0.3, the performance is the best. A smaller τ will make the quality of hard examples decrease, and too large τ will make training difficult. For the weight control ratio γ, when γ is equal to 2, the re-ID performance reaches the best. A smaller γ cause the network to focus too much on triplet loss, thus suffers from a weaker generalization capability [3]. And excessive γ will make the network lose the ability to capture the changes of the same person.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

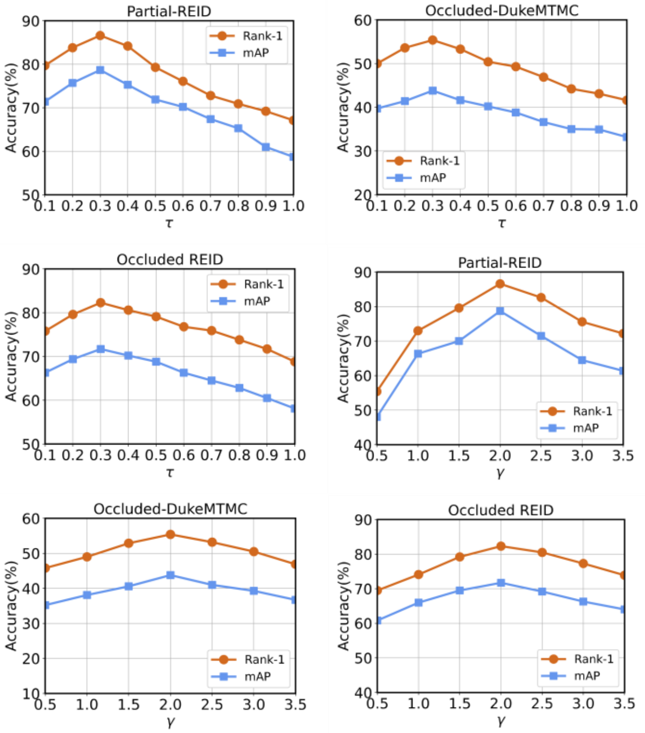

Fig.9. Rank-1 and mAP curves as function of τ and γ, when the variable is τ, γ is set to 2.0, and when γ varies, τ is set to 0.3. It can be seen that the best performance is achieved when τ=0.3 and γ=2.

TABLE XI

THE EFFECTIVENESS VALIDATION EXPERIMENTS RESULTS OF SFA, RESULTS ON PARTIAL-REID, OCCLUDED-DUKEMTMC AND OCCLUDED REID ARE REPORTED, RANK-1 ACCURACY (%) AND MAP (%) ARE SHOWN.

Methods	Partial-REID		Occluded-Duke		Occluded REID	
	R1	mAP	R1	mAP	R1	mAP
AOPS+Maskdot+VRM	79.5	73.5	48.1	36.6	73.8	63.5
AOPS+SFA+VRM	**86.8**	**78.8**	**55.4**	**43.8**	**82.5**	**71.8**

4) *Effectiveness of SFA*

In order to verify the effectiveness of SFA and to ensure the fairness of the experimental comparison, the pedestrian body masks provided by AOPS should also be used in the comparison experiment. Thus, in the comparison experiment (named "Maskdot"), we directly multiply the mask $\boldsymbol{M}$ with the feature map $\boldsymbol{H}$, so that only the features corresponding to the non-zero positions on the mask are retained. The comparison results are shown in Table XI. It can be seen that SFA brings the performance gain of +(7.3%/5.3%) in rank-1/mAP on the Partial-REID dataset. On Occluded-DukeMTMC, and Occluded REID, the gains are +(7.3%/7.2%) and +(8.5%/8.2%) in rank-1/mAP. As for the reason of this performance difference, AOPS cannot accurately segment human body parts, and occlusion information will inevitably be introduced during the test phase, which leads to a lower performance of the method "Maskdot". But as mentioned in the previous section D, SFA, combined with the attention mechanism of the AOPS masks, enables the model to obtain the ability to recognize pedestrian pictures containing background information. Therefore, SFA can use the mask to generate discriminative features better than performing a dot product directly with the mask.

F. Ablation Study on VRM

For the effectiveness of VRM, the experimental results are shown in Table XII. We can see that when VRM is used, the performance gains are +(3.5%/2.8%) on Partial-REID, +(2.5%/3.1%) on Occluded-DukeMTMC and +(2.3%/1.9%) on Occluded REID in rank-1/mAP. This is because in the case of occlusion, VRM can filter the information which is useless in the whole-body picture by replacing the mask of gallery image with that of query. It also shows that solving the problem illustrated in Fig. 1 is helpful for occluded person re-ID.

Discussion of efficiency: In order to solve the mismatching problem illustrated in Fig. 1, we proposed VRM, each time a query and a gallery image are paired to compute similarity, we replace the mask of the gallery with this query's mask to generate gallery feature tailor-made for this query. Thus, when testing on the dataset for academic research, our feature extraction times of gallery images will increase by the same multiple as the number of query pictures. However, in practical applications, the re-ID task is usually used to search for other pictures of one target of interest from the gallery pool. *i.e.*, we get a query, and we retrieve the whole gallery to search for similar items to this query. In this case, the number of feature extraction times of our method is the same as other methods. Besides, because the mask of the gallery is replaced by the mask of the query, VRM saves a large amount of gallery mask extraction time. Therefore, we recommend using VRM in more practical scenarios where there are only a few queries.

TABLE XII

THE EFFECTIVENESS VALIDATION EXPERIMENTS RESULTS OF VRM, RESULTS ON PARTIAL-REID, OCCLUDED-DUKEMTMC AND OCCLUDED REID ARE REPORTED, RANK-1 ACCURACY (%) AND MAP (%) ARE SHOWN.

Methods	Partial-REID		Occluded-DukeMTMC		Occluded REID	
	R1	mAP	R1	mAP	R1	mAP
AOPS+SFA	83.2	76.0	54.0	42.6	80.3	70.1
AOPS+SFA+VRM	**86.8**	**78.8**	**55.4**	**43.8**	**82.5**	**71.8**

G. Occluded Vehicle Re-ID

In order to further verify the generalization performance of our method, we conduct experiments on a task which is similar to the person re-ID, the vehicle re-ID. We use the VeRi-776 dataset [30]. In VeRi-776, in addition to the vehicle ID annotations, for our method, the other two important labels are the color and vehicle-model annotations (as the attributes in AOPS) provided in this dataset. In the vehicle re-ID experiment, similar to the practice of using the Market-1501 attribute to train AOPS in the person re-ID task, we use the vehicle-model and color annotations in VeRi-776 for training the AOPS. Since there is no research related to occluded vehicle re-ID, and the VeRi-776 dataset is not a dataset specifically used to evaluate the method of occlusion re-ID. We follow the setting of the person re-ID dataset Partial-REID, select pictures from the test set of VeRi-776, and perform manual color blocks like in the way shown in Fig. 7 to the query images. Thus, we can simulate occlusions on the vehicle dataset to test the effectiveness of our method.

Fig. 10 is the pedestrian and vehicle masks obtained by the pedestrian AOPS and the vehicle AOPS when there is occlusion. The value in the mask is actually binary, here we replace the white area with the original image content at the corresponding

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

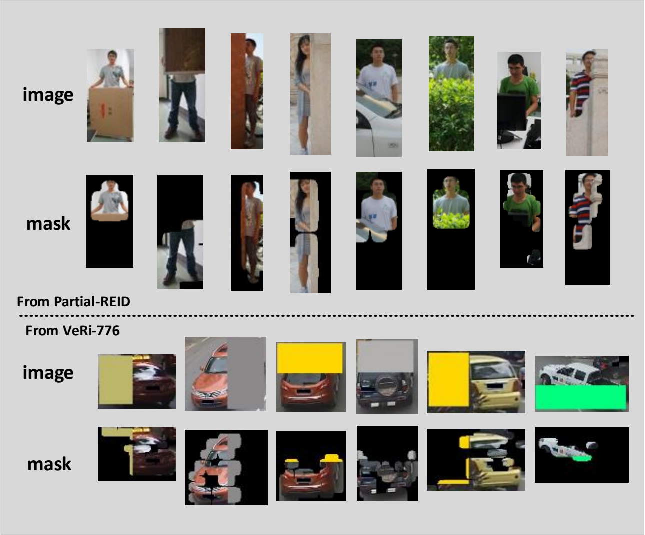

Fig. 10. The human body masks and the vehicle masks obtained by pedestrian AOPS and vehicle AOPS. Just like filtering out natural occlusions on the Partial REID dataset, AOPS trained with the vehicle color and model can also effectively avoid simulated occlusion color blocks on the VeRi-776 dataset.

position, to more easily observe the segmentation of the mask. It can be seen that the AOPS trained with the vehicle color and vehicle model can still effectively avoid most of the color blocks (simulated obstructions). Besides, it should be noted that, the AOPS used on dataset Partial-REID in Fig. 10 is trained on Market-1501, and the AOPS we use on all occluded datasets are also the same. From Fig. 7 and Fig. 10, we can see that when AOPS is used **across** datasets, its positioning effect on the human body under occlusion is still stable.

The ablation experiment results of each component in our method on VeRi-776 are shown in Table XIII. On the vehicle dataset, our method has significant improvement over the baseline. Specifically, compared to using AttrNet as the mask generator, AOPS brings a performance increase of +(16.4% /8.7%) in rank-1/mAP. And SFA achieves a performance gain of +(8.2%/4.6%) compared to direct multiplying masks. Finally, our method AOPS+SFA+VRM achieves a performance gain of +(18.3%/9.8%) over baseline. This shows that our method also has good generalization ability between different tasks.

TABLE XIII
THE EFFECTIVENESS VALIDATION EXPERIMENTS RESULTS OF EACH MODULE ON VERI-776, RANK-1 ACCURACY (%) AND MAP (%) ARE SHOWN.

Methods	VeRi-776			
	R1	R5	R10	mAP
Baseline	51.5	74.8	87.8	43.5
AttrNet+SFA+VRM	52.4	76.0	89.7	44.6
AOPS+Maskdot+VRM	60.6	82.4	92.4	48.7
AOPS+SFA+VRM	**68.8**	**90.1**	**96.8**	**53.3**

H. Visualization

Fig. 11 shows some retrieval examples of the baseline and the proposed framework (AOPS+SFA+VRM) on Partial-REID and VeRi-776. We can see that in Fig. 11 (a), on the one hand, the baseline network cannot focus on the visible (unblocked) body parts of pedestrians, thus retrieving erroneous results which have similar appearance to the target person (also can be seen in Fig. 11 (b)-(c)). On the other hand, the baseline regards grass as a part of the human body, and thus retrieves the wrong candidate containing grass in the background. In Fig. 11 (b) and (c), the persons' lower body is blocked by a table or a car. In the search results of the baseline, there are wrong candidates whose lower body texture patterns are similar to the obstacles. Similar situations with the pedestrian dataset also appear in the vehicle results in Fig. 11 (e)-(f). However, compared with the baseline, our method can focus on the part of the unblocked target of the query, and focus on the same position of the gallery during retrieval, so that the task of occluded re-ID can be completed excellently.

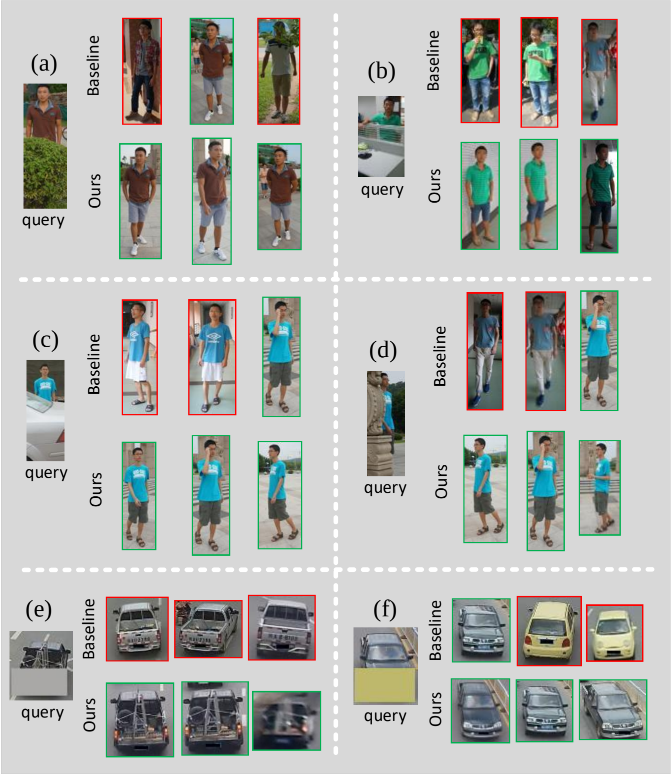

Fig. 11. Retrieval results comparison of the baseline and our method. Green and red rectangles indicate correct and error results, respectively. Fig. (a)–(d) are the results on person datasets Partial-REID and Fig. (e)-(f) are results on vehicle datasets VeRi-776. We can see that the baseline often regards the occlusions in the picture as the texture of the human body, thus searching for wrong results. Our method can effectively filter out obstructions and retrieve the correct results with valuable information in the picture.

VII. CONCLUSION AND FUTURE PLAN

Conclusion: In this paper, to solve the mismatching problem in the retrieval process of occluded person re-ID, we propose an attribute-based shift attention network (ASAN). In this framework, an attribute-guided occlusion-sensitive pedestrian segmentation (AOPS) module is trained in a weakly supervised manner to localize the human body parts in the occluded pedestrian image. Then, the localization mask and the corresponding image are input to a proposed shift feature adaption (SFA) network, by which the inference from obstructions can be effectively eliminated. Then a visible region matching (VRM) is presented to eliminate the useless information in holistic image and purify the features. Consequently, discriminative feature description is obtained.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

The experimental results on three occluded/partial pedestrian re-ID datasets show that our method outperforms existing state-of-the-art approach. The experiments on two holistic person re-ID datasets and a vehicle re-ID dataset also verify its generality.

Future Plan: Although we proposed the ASAN and achieved the state-of-the-art in this work, the most of the existing occluded person re-ID research set the occluded pedestrian images as query, and set the holistic pedestrian images as gallery. Our work was also based on this setting. In addition, the problem of people occluding each other is also widely present in real scenes and occluded datasets. The existing methods cannot well solve the problems of people occlusion and people being occluded by obstacles at the same time. In future work, we may consider the case where both query and gallery are blocked, and design methods to solve the problem of pedestrians being blocked by obstacles and being blocked by non-target pedestrians at the same time. Besides, in view of the lack of train set of occluded re-ID, the cross-domain research on occluded re-ID is also an explorable research direction based on the related ideas of cross-domain person re-ID research [63]-[64].

References

[1] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. ICCVW*, 2017, pp. 2590-2600.

[2] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, 2014, pp. 152-159.

[3] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CORR*, 2017. [Online]. Available: https://arxiv.org/abs/1703.07737.

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR* 2016, pp. 2921-2929.

[5] L. He, W. Liu, "Guided Saliency Feature Learning for Person Re-identification in Crowded Scenes," in *Proc. ECCV*, 2020, pp. 357-373.

[6] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute Attention Network for Person Re-Identifications," in *Proc. CVPR*, 2019, pp. 7134-7143.

[7] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J.-H. Lai, and S. Gong, "Partial person re-identification," in *Proc. ICCV*, 2015, pp. 4678-4686.

[8] L. He, J. Liang, H. Li and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. CVPR*, 2018, pp. 7073-7082.

[9] L. He, Z. Sun, Y. Zhu, and Y. Wang, "Recognizing partial biometric patterns," *CORR*, 2018. [Online]. Available: https://arxiv.org/abs/1810.07399.

[10] H. Luo, X. Fan, C. Zhang, and W. Jiang, "STNReID: Deep Convolutional Networks with Pairwise Spatial Transformer Networks for Partial Person Re-identification," *CORR*, 2019. [Online]. Available: https://arxiv.org/abs/1903.07072.

[11] 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. CVPR*, 2019, pp. 393-402.

[12] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-Guided Feature Alignment for Occluded Person Re-Identification," in *Proc. ICCV*, 2019, pp. 542-551.

[13] J. Zhou, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. ICME*, 2018, pp. 1-6.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis*, vol. 128, no. 2, pp. 336-359, 2020.

[15] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *Proc. WACV*, 2018, pp. 839-847.

[16] H.-M. 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.

[17] 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.

[18] S. Bak, P. Carr, "Deep Deformable Patch Metric Learning for Person Re-Identification," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 28, no. 10, pp. 2690-2702, 2018.

[19] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng., "Deep Hybrid Similarity Learning for Person Re-Identification," *IEEE Trans. Circuits Syst. Video Technol*. vol. 28, no. 11, pp. 3183-3193, 2018.

[20] 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 *Proc. ICCV*, 2019, pp. 8449-8458.

[21] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR*, 2011, pp. 649-656.

[22] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, 2015, pp. 1116-1124.

[23] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proc. ICCV*, 2017, pp. 3774-3782.

[24] A. Verma, H. Qassim, and D. Feinzimer, "Residual squeeze CNDS deep learning CNN model for very large scale places image recognition," in *Proc UEMCON*, 2017, pp. 463-469.

[25] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. ICLR*, 2015.

[26] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," in *Proc. NIPS*, 2015, pp. 2017-2025.

[27] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment," in Proc. ICCV, 2019, pp. 3622-3631.

[28] G. Wang, Y. Yuan, X. Chen, J. Li, and Xi Zhou, "Learning Discriminative Features with Multiple Granularities for Person Re-Identification," in *Proc. ACM MM*, 2018, pp. 274-282.

[29] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human Semantic Parsing for Person Re-Identification," in *Proc. CVPR*, 2018, pp. 1062-1071.

[30] X. Liu, W. Liu, T. Mei, and H. Ma, "A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance," in *Proc. ECCV*, 2016, pp. 869-884.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[32] L. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," in *Proc. ICLR*, 2017.

[33] 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. CVPR*, 2020.

[34] R. Layne, T. M. H., S. Gong, "Re-id: Hunting Attributes in the Wild," in *Proc. BMVC*, 2014.

[35] Y. Yang, J. Yang, J. Yan, S.Liao, D. Yi, and S. Z. Li, "Salient Color Names for Person Re-identification," in *Proc. ECCV*, 2014, pp. 536-551.

[36] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S. Davis, and Wen Gao, "Multi-Task Learning with Low Rank Attribute Embedding for Multi-Camera Person Re-Identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1167-1181, 2018.

[37] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint Learning for Attribute-Consistent Person Re-Identification," in *Proc. ECCVW*, 2014, pp. 134-146.

[38] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proc. ICPR*, 2016, pp. 2428-2433.

[39] Y. Chen, S. Duffner, A. Stoian, J.-Y.Dufour, and A. Baskurt, "Deep and low-level feature based attribute learning for person re-identification," *Image Vis. Comput.*, vol. 79, pp. 25-34, 2018.

[40] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A Richly Annotated Dataset for Pedestrian Attribute Recognition," *CORR*, 2016. [Online]. Available: https://arxiv.org/abs/1603.07054.

[41] Y. Deng, P.Luo, C. C.Loy, and X. Tang, "Pedestrian Attribute Recognition At Far Distance," in *Proc. ACM Multimedia*, 2014, pp. 789-792.

[42] D. Tao, L. Jin, Y. Wang, Y. Yuan, X. Li, "Person Re-Identification by Regularized Smoothing KISS Metric Learning," *IEEE Trans. Circuits Syst. Video Technol*. vol. 23, no. 10, pp. 1675-1685, 2013.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3088446, IEEE Transactions on Circuits and Systems for Video Technology

[43] P. Zhang, J. Lai, Q. Zhang, and X. Xie, “MGD: Mask Guided De-occlusion Framework for Occluded Person Re-identification,” in *Proc. ISciDE*, 2019, pp. 411-423.

[44] 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. ECCV*, 2018, pp. 501-518.

[45] K. Wang, P. Wang, C. Ding, and D. Tao, “Batch Coherence-Driven Network for Part-Aware Person Re-Identification,” *IEEE Trans. Image Process*. vol. 30, pp. 3405-3418, 2021.

[46] Y. Li, H. Yao, T. Zhang, and C. Xu, “Part-based Structured Representation Learning for Person Re-identification,” *ACM Trans. Multim. Comput. Commun. Appl.* vol. 16, no. 4, pp. 134:1-134:22, 2021.

[47] 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.

[48] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, “Devil in the Details: Towards Accurate Single and Multiple Human Parsing,” in *Proc. AAAI*, 2019, pp. 4814-4821.

[49] X. Liu, W. Liu, T. Mei, and H. Ma, “PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance,” *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645-658, 2018.

[50] J. Miao, Y. Wu and Y. Yang, "Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification," in I*EEE Trans. Neural Networks Learn. Syst.* doi: 10.1109/TNNLS.2021.3059515.

[51] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised Person Re-identification: Clustering and Fine-tuning,” in *ACM Trans. Multim. Comput. Commun. Appl.* Vol. 14, no. 4, pp. 83:1-83:18, 2018.

[52] H. Fan, X. Chang, D. Cheng, Y. Yang, D. Xu, and A. G. Hauptmann, “Complex Event Detection by Identifying Reliable Shots from Untrimmed Videos,” in *Proc. ICCV*, 2017, pp. 736-744.

[53] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, “Auto-ReID: Searching for a Part-Aware ConvNet for Person Re-Identification,” in *Proc. ICCV*, 2019, pp. 3749-3758.

[54] Fan H, Liu P, Xu M, and Yang Y. “Unsupervised Visual Representation Learning via Dual-Level Progressive Similar Instance Selection,” *IEEE Trans Cybern.* 2021 Mar 11;PP. doi: 10.1109/TCYB.2021.3054978.

[55] Y. Huang, S. Lian, S. Zhang, H. Hu, D. Chen, and T. Su, “Three-Dimension Transmissible Attention Network for Person Re-Identification,” *IEEE Trans. Circuits Syst. Video Technol.* vol. 30, no. 12, pp. 4540-4553, 2020.

[56] H. Li, S. Yan, Z. Yu, and D. Tao, “Attribute-Identity Embedding and Self-Supervised Learning for Scalable Person Re-Identification,” *IEEE Trans. Circuits Syst. Video Technol*. vol. 30, no. 10, pp. 3472-3485, 2020.

[57] S. Ge, C. Li, S. Zhao, and D. Zeng, “Occluded Face Recognition in the Wild by Identity-Diversity Inpainting,” *IEEE Trans. Circuits Syst. Video Technol.* vol. 30, no. 10, pp. 3387-3397, 2020.

[58] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, “SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification,” in *Proc. ICCV*, 2019, pp. 9526-9535.

[59] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, and Z. Zhang, “Beyond Scalar Neuron: Adopting Vector-Neuron Capsules for Long-Term Person Re-Identification,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3459-3471, 2020.

[60] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, “Learning Part-based Convolutional Features for Person Re-Identification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 902-917, 2021.

[61] W. Wu, D. Tao, H. Li, Z. Yang, and J. Cheng, “Deep features for person re-identification on metric learning,” *Pattern Recognit*., vol. 110, pp. 107424, 2021.

[62] L. Zhang, F. Liu, and D. Zhang, “Adversarial View Confusion Feature Learning for Person Re-Identification,” *IEEE Trans. Circuits Syst. Video Technol.* vol. 31, no. 4, pp 1490-1502, 2021.

[63] H. Li, Y. Chen, D. Tao, Z. Yu, and G. Qi, “Attribute-Aligned Domain-Invariant Feature Learning for Unsupervised Domain Adaptation Person Re-Identification,” *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1480-1494, 2021.

[64] Y. Huang, Q. Wu, J. Xu, Y. Zhong, “SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification,” in *Proc. ICCV*, 2019, pp. 9526-9535.

[65] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, Z. Zhang, “Beyond Scalar Neuron: Adopting Vector-Neuron Capsules for Long-Term Person Re-Identification,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3459-3471, 2020.

Hanyang Jin received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2014, and M.S. degree from Guilin University of Electronic Technology, Guilin, China, in 2017. He is currently pursing the Ph.D. degree with Xi’an Jiaotong University. His research interests are computer vision and multimedia retrieval.

Shenqi Lai received the B.S. degree from Xi’an Jiaotong University, Xi’an, China, in 2014, where he received the M.S. degree from the School of Software Engineering. His research interests are multimedia retrieval, neural network acceleration and computational aesthetics.

Xueming Qian (M’10) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in electronics and information engineering from Xi’an Jiaotong University, Xi’an, China, in 2008. He was a Visiting Scholar with Microsoft Research Asia, Beijing, China, from 2010 to 2011. He was previously an Assistant Professor at Xi’an Jiaotong University, where he was an Associate Professor from 2011 to 2014, and is currently a Full Professor. He is also the Director of the Smiles Laboratory, Xi’an Jiaotong University. His rsearch interests include social media big data mining and search.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Xian Jiaotong University. Downloaded on July 26,2021 at 03:16:20 UTC from IEEE Xplore. Restrictions apply.