# Occluded Person Re-Identification via Defending Against Attacks From Obstacles

Shujuan Wang, Run Liu, Huafeng Li, Guanqiu Qi, and Zhengtao Yu

***Abstract*— Due to incomplete appearance features, the identity matching of occluded pedestrians under multiple cross-camera views is a long-term challenge. Although existing re-identification (re-ID) solutions of occluded pedestrians have made significant progress, most of them achieve accurate identity matching by extracting pedestrian appearance features from unoccluded areas. However, when a pedestrian is partially blocked by the body of another pedestrian, existing methods cannot accurately determine whether the unoccluded body parts belong to the target pedestrian, which brings great difficulties to pedestrian identity matching. To alleviate this problem, this paper introduces the idea of adversarial attack into occluded person re-ID and proposes an adversarial training framework that can defend against attacks from obstacles to resist the interference of obstacles on pedestrian identity matching. Unlike existing solutions, the proposed framework is not limited to extracting features of unoccluded human body areas to achieve occluded person re-ID, but explores how to make the re-ID model more resistant to obstacles. In the proposed framework, the occluded pedestrian images are regarded as adversarial examples and used to attack model training. If the trained model can defend against this kind of attack, its generalization is significantly improved, and the above-mentioned issues are also effectively solved. Specifically, a single-branch dual-stream collaborative network is designed. With the cooperation of the pre-trained verification guidance network, the model realizes the attack and defense of adversarial samples. This work broadens research horizons in robust model design of occluded person re-ID, and expands the scope of adversarial attacks. Compared with existing solutions, a lot of experimental results confirm that the proposed solution achieves better performance on two occluded re-ID datasets and two partial re-ID datasets.**

***Index Terms*— Occluded person re-ID, adversarial examples, adversarial attack.**

Manuscript received 13 January 2022; revised 19 June 2022; accepted 17 October 2022. Date of publication 31 October 2022; date of current version 7 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62276120, Grant 61966021, and Grant 61562053; in part by the National Key Research and Development Plan Project under Grant 2018YFC0830105 and Grant 2018YFC0830100; and in part by the Yunnan Natural Science Funds under Grant 2018FY001(-013), Grant 2016FB105, Grant 2017FB094, Grant 2016FD039, and Grant 2016FB109. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. William R. Schwartz. *(Shujuan Wang and Run Liu contributed equally to this work.) (Corresponding authors: Huafeng Li; Guanqiu Qi.)*

Shujuan Wang, Run Liu, Huafeng Li, and Zhengtao Yu are with the Yunnan Provincial Key Laboratory of Artificial Intelligence, Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China (e-mail: shujuanwang0703@126.com; liurun_531@163.com; lhfchina99@kust.edu.cn; ztyu@hotmail.com).

Guanqiu Qi is with the Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222 USA (e-mail: qig@buffalostate.edu).

Data is available online at https://github.com/lhf12278/OPR-DAAO.

Digital Object Identifier 10.1109/TIFS.2022.3218449

## I. Introduction

PERSON re-identification (re-ID) is used to determine whether pedestrian images captured by non-overlapping cameras contain specific pedestrians. Since person re-ID related solutions can provide key technical support for finding criminal suspects and missing people, they have received extensive attention from researchers [1], [2], [3], [4], [5], [6], [7], [8], [9]. Although significant research progress has been made in person re-ID in recent years, pedestrian appearance features are easily interfered by obstacles, which causes a huge challenge on the popularization and application of existing technologies. Therefore, occluded person re-ID solutions have been proposed to alleviate the above-mentioned issues [10], [11].

Although existing occluded person re-ID solutions can effectively alleviate the negative impact of obstacles on re-ID performance, they mainly focus on extracting the features from unoccluded areas to achieve person identity matching. As a key problem, most of these existing solutions need to detect or predict unoccluded areas in pedestrian images. To this end, various feature extraction methods based on external models (such as pedestrian key point detection models [12] and human semantic parsing models [13], [14]) have been proposed for occluded pedestrians. Particularly, Kalayeh et al. [15] proposed an occluded area detection method based on the human semantic parsing model. Miao et al. [11] applied a pedestrian pose estimation model to indicate occluded areas. Gao et al. [16] proposed a pose-guided visibility predictor to estimate whether various human body parts are occluded. Wang et al. [17] used a key point detection model to construct an adaptive direction graph convolutional layer to achieve the prediction of occluded areas, and the message passing of meaningless features was automatically suppressed by dynamically learning both direction and degree of linkage, therefore suppressing the interference of obstacles. Zhang et al. [18] inserted a semantic branch into the overall framework of occluded person re-ID to generate foreground-background masks of pedestrian images for the detection of unoccluded areas. If the corresponding obstacle is not part of human body, external model-based methods tend to have better performance when a target pedestrian is occluded.

However, in addition to surrounding objects that cause the obstruction of pedestrians in varying degrees, pedestrians are often occluded by parts of other pedestrians in real-world scenes. In this case, existing person re-ID models are difficult to distinguish the target pedestrians from all pedestrians. So,

(a) (b)

Fig. 1. External models show low robustness when target pedestrians are occluded by other pedestrians. The sub-images (a) and (b) show the performance of a key point detection model [12] and a human semantic parsing model [13], [14] respectively, when target pedestrians are occluded by other pedestrians.

external models used in existing solutions always return false detection results, thereby limiting the further performance improvement of these solutions. As shown in Fig.1, when target pedestrians are occluded by parts of other pedestrians, both the key point detection model and human semantic parsing model return incorrect detection results. In this case, the person re-ID models designed based on these external models may introduce features that are inconsistent with the target pedestrian identities. Therefore, errors occur in pedestrian identity matching, thereby reducing overall model performance.

To alleviate the above-mentioned problems, this paper designs one occluded person re-ID solution from a brand-new view. Specifically, it propose an occluded person re-ID framework with integrated adversarial attacks. Unlike existing methods, the proposed framework is no longer dedicated to detecting non-occluded human body parts to achieve occluded person re-ID. Instead, in the proposed method, obstacles are treated as disturbances added to non-occluded pedestrian images, while occluded pedestrian samples are treated as adversarial examples. Correspondingly, occluded person re-ID is viewed as a defense against attacks from obstacles. If the re-ID model can defend the attacks of adversarial samples, its generalization ability in occluded pedestrian data is significantly improved. For the first time, the idea of adversarial attack and defense is introduced into occluded person re-ID to solve the interference of obstacles on pedestrian identity matching. Since there is no need to use any external model (such as key point detection model, human semantic parsing model) to assist re-ID model training, the negative impact of external model detection results on the re-ID model can be effectively avoided.

Technically, this paper proposes a verification guidance adversarial training framework to resist attacks from obstacles. It is mainly composed of adversarial sample synthesis, verification guidance network, and single-branch dual-stream collaborative network (SDCN). The adversarial sample synthesis has three main steps, adding obstacles to full pedestrian images, generating occluded pedestrian images, and composing fully-occluded pedestrian image pairs with the original unoccluded pedestrian images. The composed image pairs are used to train both verification guidance network and SDCN. The verification guidance network mainly plays a role in guiding and supervising SDCN training, which ensures SDCN has a strong ability to defend against attacks from obstacles. The features from the feature encoder of the verification guidance network are divided into two directional streams. One is used to receive the features of occluded pedestrians extracted by SDCN. The other is used to receive the features of full pedestrian images extracted by SDCN. After the features extracted by the verification guidance network are synthesized with the dual-stream features from SDCN, the fused features are input to a classifier in the verification guidance network for verification, and the optimization of SDCN parameters is reversely guided according to the verification results.

As the core idea of this design, if SDCN can extract features that are not affected by obstacles from occluded pedestrian images, these features should be compatible with the corresponding full pedestrian image features from the verification guidance network. After the synthesis, the synthesized features should still be correctly classified by the pre-trained classifier of the verification guidance network. For the network structure, SDCN only contains one network branch, but there are two input streams. One input stream is used to receive the full pedestrian images, and the other input stream is used to receive the synthesized occlusion images corresponding to full pedestrian images. The first input stream ensures model performance on full pedestrian images. The adversarial training of the re-ID model is implemented in the second input stream to enable the model to defend against the attacks from adversarial examples. In this process, the dual-stream input is applied to the cooperative training of SDCN to improve the robustness of SDCN to obstacles under the supervision of the pre-trained verification guidance network.

This paper has three main contributions as follows.

- This paper introduces the idea of adversarial attack and defense into occluded person re-ID for the first time, treating obstacles in occluded samples as disturbances added to the corresponding unoccluded pedestrian images. Correspondingly, occluded person samples can be regarded as adversarial samples. Occluded person re-ID is achieved by defending against attacks from obstacles. This idea not only broadens the horizon of adversarial attack and defense, but also opens up a new path for the design of occluded person re-ID solutions, which is expected to inspire more new ideas on occluded person re-ID.
- An adversarial defense framework is designed for defending against attacks from occluded samples. This framework consists of two main parts, SDCN and verification guidance network. Under the guidance and cooperation of the pre-trained verification guidance network, this framework realizes SDCN adversarial training

on both full pedestrian samples and synthesized occlusion samples collaboratively. Correspondingly, SDCN gains the ability to defend against attacks from adversarial samples (occluded samples). Therefore, the robustness of the re-ID model to obstacles is improved.

- The proposed method does not need even any assistance from an external model (such as key point detection model, human semantic parsing model). The comparative experimental results confirm that the proposed method outperforms state-of-the-art methods on occluded person re-ID. Additionally, the proposed method also shows effectiveness on both partial person re-ID and general person re-ID.

The rest of this paper is organized as follows. Section II discusses related work; Section III elaborates the proposed method; Section IV analyzes the comparative experimental results; and Section V concludes this paper.

## II. Related Work

### A. Occluded Person Re-ID

Pedestrians are inevitable to be occluded in real-world scenes. Therefore, it is necessary to consider the impact of obstacles on the performance of person re-ID methods. Zhuo et al. [10] first proposed the concept of occluded person re-ID and designed a human body attention framework to extract the appearance features of occluded pedestrians. Subsequently, occluded person re-ID has attracted the attention of researchers, and a large number of effective methods have been proposed. According to the type of feature extraction, existing occluded person re-ID methods can be categorized into external model-assisted methods and external model-free methods. The methods based on external model assistance mainly include methods based on pedestrian pose estimation models and methods based on human semantic parsing models. In particular, Miao et al. [11] applied a pedestrian pose estimation model to extract the key points of occluded pedestrians. According to the unoccluded key points, the extraction and enhancement of unoccluded features were realized. Gao et al. [16] proposed a pose guided visibility predictor to estimate whether various parts of the body are occluded. Wang et al. [17] proposed to use the corresponding features of pedestrian key points as the node features of the graph structure and automatically suppress the interference of obstructions by dynamically learning both direction and degree of linkage between graph nodes. Zhang et al. [18] integrated a human body parsing branch into the overall framework of occluded person re-ID to realize the semantic segmentation of the human body. The foreground-background masks of pedestrian images were generated to detect the unoccluded areas. Ma et al. [19] proposed a gesture-guided inter- and intra-part relational transformer to solve the issues of occluded person re-ID, and introduced a transformer to establish part-aware long-term correlations.

Although the above-mentioned methods based on external models can obtain excellent re-ID performance, the performance of these methods relies heavily on the detection results of external models. When a pedestrian is occluded by another pedestrian, two pedestrians are likely to appear in one image. In this case, external models are likely to return wrong detection results, which reduce the corresponding model performance. To alleviate this issue, a method [20] was proposed to use pose estimation only in the training process to regularize the learning of semantic alignment features. Although this can reduce the re-ID model's dependence on pose estimation, it cannot get rid of the dependence completely.

As an external model-free method, Zhuo et al. [21] equipped a co-saliency network in the teacher-student network to detect unoccluded areas. Although this method does not introduce an external model, it needs to detect human body areas in the feature extraction process, which reduces its robustness in dense pedestrian scenes. To extract pedestrian features from unoccluded areas, He et al. [22] proposed foreground perception pyramid reconstruction (FPR) without feature alignment to accurately calculate the matching scores between occluded pedestrians. Li et al. [23] proposed a Transformer that senses human body parts to solve the issues of occluded person re-ID. Although external model-free methods completely get rid of the constraints of external models, they still face a huge challenge in the perception of the areas where pedestrians are not occluded. In order to alleviate this issue, this paper converts the problem of occluded person re-ID into a problem of defending the adversarial attacks of obstacles. The performance of occluded person re-ID is enhanced by improving the model's ability to defend adversarial attacks. So the challenge from pedestrians occluded by other pedestrians can be alleviated.

### B. Partial Person Re-ID

In real-world scenes, to avoid the impact of obstacles on pedestrian identity matching, obstacles in the query images are usually segmented from the corresponding pedestrian images, and the remaining unoccluded parts of pedestrian images (partial person) are used in pedestrian identity matching, which is called partial person re-ID. Zheng et al. [24] proposed a global-to-local matching model to achieve partial person re-ID. He et al. [25] proposed depth space feature reconstruction to avoid the explicit alignment of features in partial person re-ID. Sun et al. [26] proposed a visibility-aware part model to solve extreme spatial dislocation, when directly comparing partial pedestrian images with the overall pedestrian images in partial person re-ID. Luo et al. [27] proposed a deep partial person re-ID framework based on pairwise spatial transformer networks to solve the matching issue between partial person images and holistic person images. Unlike existing partial person re-ID methods, the proposed method focuses on improving the model's robustness in the presence of obstacles. Even so, the proposed method can still be used on partial person re-ID, and show excellent re-ID performance.

### C. Adversarial Attacks in Person Re-ID

The concept of adversarial attacks was first proposed by Szegedy et al. [28]. As the main purpose, it generates

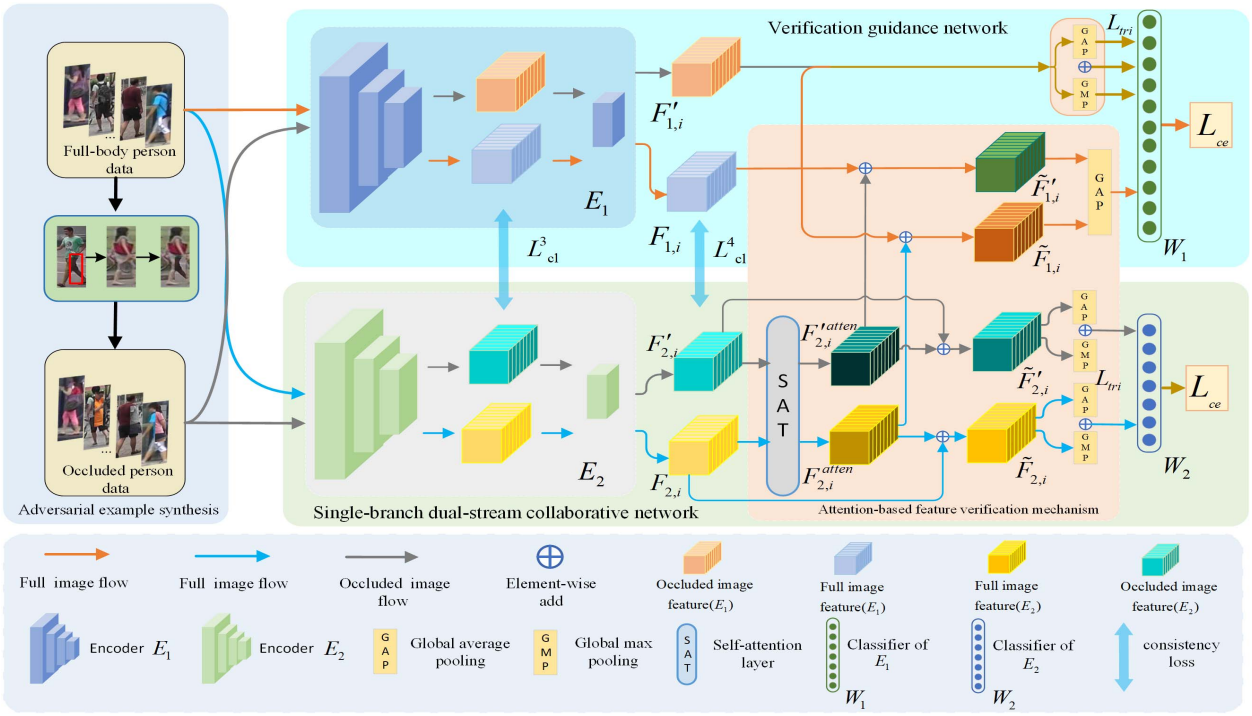



Fig. 2. Overview of the proposed framework. It consists of occluded sample synthesis, verification guidance network, and SDCN. In the occluded sample synthesis, occluded pedestrian samples required by model training are generated, and used as adversarial examples to participate in the training of both verification guidance network and SDCN. Both verification guidance network and SDCN are trained under occluded and full pedestrian image samples, respectively. After the verification guidance network training is completed, both occluded and full pedestrian images are input into SDCN in pairs. Additionally, the full pedestrian images are input into the pre-trained verification guidance network. The adversarial training of SDCN is completed under the constraints and verification of the verification guidance network.

adversarial examples misclassified by the deep neural network (DNN)-based model by adding small perturbation magnitudes to the original inputs. For adversarial attacks in person re-ID, most of existing methods focus on generating adversarial examples to invalidate DNN-based person re-ID models. Wang et al. [29] first explored whether the deep-based person re-ID system is susceptible to adversarial attacks, and proposed a method called advPattern to generate adversarial examples, which was applied to generate adversarial patterns on pedestrian clothes. To explore whether deep-based person re-ID systems are vulnerable to adversarial attacks, Wang et al. [30] developed a multi-level network architecture to perform back-box attacks and proposed a perceptual loss to ensure that the attacks are not conspicuous. Wang et al. [31] proposed a multi-expert adversarial attack detection method, which achieved the detection of adversarial attacks by checking inconsistency in contexts. In order to alleviate this shortcoming of DNN, many researchers proposed to generate adversarial examples and apply them to model training, so the corresponding models gained the ability to defend these perturbations. This process is called adversarial training [32]. Based on this idea, this paper treats obstacles as a type of perturbation. It is the first time to introduce adversarial attacks into occluded person re-ID. A person re-ID model is proposed to defend the interference from obstacles. Unlike existing adversarial training, the perturbation caused by obstacles is no longer small in the proposed method. The proposed method realizes occluded person re-ID by improving its robustness to obstcles, and avoids the challenges in detecting or determining unoccluded pedestrian areas.

## III. The Proposed Approach

### A. Overview

As shown in Fig.2, the proposed method is mainly composed of adversarial example (i.e. occluded pedestrian images) synthesis, verification guidance network, and SDCN. In adversarial sample synthesis, occluded images are generated based on unoccluded pedestrian images, and then used as adversarial samples to improve the model's robustness to obstacles. The verification guidance network is mainly composed of an encoder $\boldsymbol{E}_1$ and a classifier $\boldsymbol{W}_1$. This network is first pre-trained on both occluded and full pedestrian samples, and then used to guide SDCN training. It determines whether SDCN is robust to occluded pedestrian images and adjusts SDCN training according to the corresponding result. SDCN is mainly composed of an encoder $\boldsymbol{E}_2$, a self-attention layer, and a classifier $\boldsymbol{W}_2$. The encoder $\boldsymbol{E}_2$ is mainly used to extract input image features. The self-attention layer is mainly used to enable the network to focus on more discriminative areas and suppress unwanted features. In the training process, the network receives two data streams from both occluded and full pedestrian images. The cooperative training of SDCN is realized on dual-stream data under the guidance of the verification network. Therefore, the trained SDCN gains a strong ability to defend the attacks from obstacles.

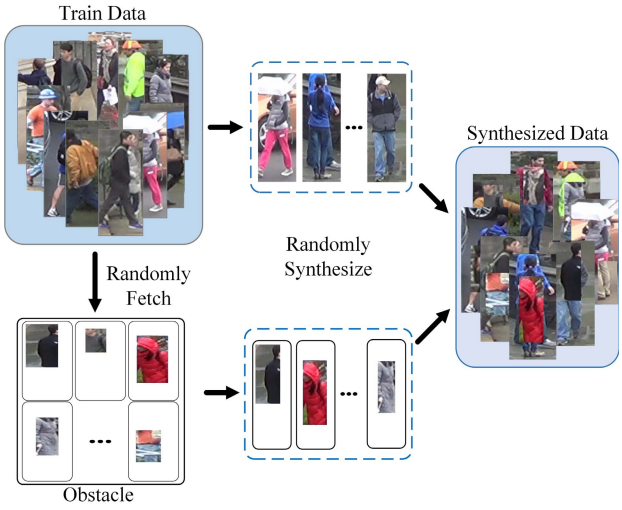



Fig. 3. Overview of adversarial example synthesis.

### B. Synthesis of Adversarial Samples

The synthesis of adversarial example is shown in Fig.3. Specifically, given an unoccluded training set $\boldsymbol{D}_t = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_t}$, where $N_t$ is the total number of images in $\boldsymbol{D}_t$, $\boldsymbol{x}_i$ is the $i$-th image, $\boldsymbol{y}_i \in \{1, 2, \cdots, L_t\}$ is the identity label corresponding to $\boldsymbol{x}_i$, and $L_t$ is the total number of pedestrian identities. In order to obtain adversarial examples and realize the dual-stream collaborative network training, a random patch method is applied to generate occluded pedestrian images (adversarial examples) on the unoccluded pedestrian images from $\boldsymbol{D}_t$. This ensures that each generated occluded pedestrian image has a corresponding unoccluded pedestrian image. Specifically, for any image $\boldsymbol{x}_i \in \boldsymbol{D}_t$, another image $\boldsymbol{x}_j (i \neq j)$ is randomly selected from $\boldsymbol{D}_t$ as the obstacle source. A rectangular area of the image $\boldsymbol{x}_j$ is randomly cropped, and the image contents of this area are pasted to the same position of the image $\boldsymbol{x}_i$ to obtain the occluded image $\boldsymbol{x}'_i$. The dataset formed by the occluded image $\boldsymbol{x}'_i$ is denoted as $\boldsymbol{D}'_t = \{\boldsymbol{x}'_i, \boldsymbol{y}_i\}_{i=1}^{N_t}$. $(\boldsymbol{x}'_i, \boldsymbol{x}_i)$ is a sample pair composed of a full pedestrian sample and an occluded pedestrian sample.

Suppose the images in the training sample set $\boldsymbol{D}_t$ are all resized to $W \times H$. The size of the rectangle randomly cropped from $\boldsymbol{x}_j$ is $W_a \times H_a$. This rectangle is used as the obstacle at the corresponding position of $\boldsymbol{x}_i$. $W_a \times H_a / W \times H \in [s_l, s_h]$, and $s_l$, $s_h$ are the lower and upper bounds allowed by $W_a \times H_a / W \times H$, respectively. Suppose $W_a / H_a = r_e$, $S_a = W_a \times H_a$, the length and width of the cropped area from $\boldsymbol{x}_j$ can be expressed as $H_a = \sqrt{S_a / r_e}$, $W_a = \sqrt{S_a \times r_e}$, respectively. In order to ensure that the cropped area does not fall outside the image $\boldsymbol{x}_j$, the coordinates of the top left vertex of the cropped rectangle area are defined as $(\boldsymbol{x}_a, \boldsymbol{y}_a)$, which satisfies $\boldsymbol{x}_a + W_a \leq W$ and $\boldsymbol{y}_a - H_a \geq 0$. Under this condition, the rectangular area enclosed by the upper left vertex coordinate $(\boldsymbol{x}_a, \boldsymbol{y}_a)$ and the lower right coordinate $(\boldsymbol{x}_a + W_a, \boldsymbol{y}_a - H_a)$ is selected as the cropped area, and this area is used to replace the area at the corresponding position of $\boldsymbol{x}_i$ to generate the occluded image $\boldsymbol{x}'_i$. The aspect ratio is set as $r_e = 0.3$, $S_l = 0.05$, $S_h = 0.2$ in this paper. The adversarial example $\boldsymbol{x}'_i$ generated by the above method is not a traditional adversarial example because the obstacles here are clearly visible. This paper expands the horizons of adversarial attacks. The above-mentioned methods are used to simulate adversarial examples, so that the idea of adversarial attack and defense is subsequently introduced to improve the re-ID model's robustness to occlusion.

### C. Verification Guidance Network

In this paper, the verification guidance network is used to assist SDCN training, so that SDCN has a strong ability to defend against obstacles. It is mainly composed of an encoder $\boldsymbol{E}_1$ and a classifier $\boldsymbol{W}_1$. The encoder $\boldsymbol{E}_1$ is mainly used to extract the features of input images. The classifier $\boldsymbol{W}_1$ is utilized to classify the input image features, and is also used to verify the features output by SDCN to determine whether SDCN extracts robust features from the input occluded images that are not affected by obstacles.

To achieve the above purpose, the full image training set $\boldsymbol{D}_t = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_t}$ and the occluded image training set $\boldsymbol{D}'_t = \{\boldsymbol{x}'_i, \boldsymbol{y}_i\}_{i=1}^{N_t}$ are used to train the verification network. Due to the occluded areas of samples in $\boldsymbol{D}'_t$ are randomly generated, and $\boldsymbol{D}'_t$ is also used to train the verification guidance network, which can encourage the network to focus on different pedestrian body parts. Therefore, it is conducive to extracting more complete appearance features of pedestrians. It assumes the training sample $\boldsymbol{x}_i$ comes from $\boldsymbol{D}_t$. The feature vectors $\boldsymbol{f}_{a,i} = GAP(\boldsymbol{E}_1(\boldsymbol{x}_i))$ and $\boldsymbol{f}_{m,i} = GMP(\boldsymbol{E}_1(\boldsymbol{x}_i))$ are extracted by the encoder $\boldsymbol{E}_1$ in the verification guidance network after being processed by global average pooling (GAP) and global maximum pooling (GMP) respectively. Similarly, for the occluded image $\boldsymbol{x}'_i$, the feature vectors are $\boldsymbol{f}'_{a,i} = GAP(\boldsymbol{E}_1(\boldsymbol{x}'_i))$ and $\boldsymbol{f}'_{m,i} = GMP(\boldsymbol{E}_1(\boldsymbol{x}'_i))$ after passing through the feature encoder $\boldsymbol{E}_1$ and the pooling layers respectively. To integrate the advantages of GAP and GMP, the integrated features $\boldsymbol{f}_{s,i} = \boldsymbol{f}_{a,i} + \boldsymbol{f}_{m,i}$ and $\boldsymbol{f}'_{s,i} = \boldsymbol{f}'_{a,i} + \boldsymbol{f}'_{m,i}$ are used as the pedestrian appearance features, and then the following cross-entropy loss and triplet loss are used to train the verification guidance network.

Specifically, the cross-entropy loss is expressed as follows:

$$\begin{aligned}\boldsymbol{L}_{ce1}(\boldsymbol{E}_1, \boldsymbol{W}_1) = -\frac{1}{n_b}\Big(&\sum_{i=1}^{n_b}\sum_{l=1}^{L_t} \boldsymbol{I}_{(l=y_i)} \log \boldsymbol{P}_l(\boldsymbol{W}_1(\bar{\boldsymbol{f}}_{a,i})) \\ &+ \sum_{i=1}^{n_b}\sum_{l=1}^{L_t} \boldsymbol{I}_{(l=y_i)} \log \boldsymbol{P}_l(\boldsymbol{W}_1(\bar{\boldsymbol{f}}_{m,i})) \\ &+ \sum_{i=1}^{n_b}\sum_{l=1}^{L_t} \boldsymbol{I}_{(l=y_i)} \log \boldsymbol{P}_l(\boldsymbol{W}_1(\bar{\boldsymbol{f}}_{s,i}))\Big), \quad (1)\end{aligned}$$

where $\bar{\boldsymbol{f}}_{k,i} = \{\boldsymbol{f}_{k,i}, \boldsymbol{f}'_{k,i}\}, k = \{a, m, s\}$, $\boldsymbol{n}_b$ is the batch size, and $\boldsymbol{I}_{(l=y_i)}$ is an indicator function, which is defined

as follows:

$$\boldsymbol{I}_{[l=y_i]} = \begin{cases} 1 - \dfrac{L_t - 1}{L_T}\varepsilon, & if\ l = y_i \\ \dfrac{\varepsilon}{L_t}, & \text{otherwise}, \end{cases} \tag{2}$$

where $\varepsilon$ is a small constant. It is set to 0.1. $\boldsymbol{P}_l$ represents the predicted logits of the identity $l$. In addition, the triplet loss is expressed as follows:

$$\begin{aligned} \boldsymbol{L}_{tri1}(\boldsymbol{E}_1, \boldsymbol{W}_1) = \frac{1}{n_b}\Big( & \sum_{i=1}^{n_b}[\max_{\tilde{i}} \bar{d}_{a,i,\tilde{i}} - \min_{j} \bar{d}_{a,i,j} + \tau_1]_+ \\ & + \sum_{i=1}^{n_b}[\max_{\tilde{i}} \bar{d}_{m,i,\tilde{i}} - \min_{j} \bar{d}_{m,i,j} + \tau_2]_+ \\ & + \sum_{i=1}^{n_b}[\max_{\tilde{i}} \bar{d}_{s,i,\tilde{i}} - \min_{j} \bar{d}_{s,i,j} + \tau_3]_+\Big), \end{aligned} \tag{3}$$

where $[z]_+ = \max\{z, 0\}$, $\bar{d}_{k,i,\tilde{i}} = d_{k,i,\tilde{i}}$, $d'_{k,i,\tilde{i}}$ $(k = a, m, s)$; $d_{k,i,\tilde{i}}$ $(d'_{k,i,\tilde{i}})$ is the Euclidean distance between $\boldsymbol{f}_{k,i}$ $(\boldsymbol{f}'_{k,i})$ and $\boldsymbol{f}_{k,\tilde{i}}$ $(\boldsymbol{f}'_{k,\tilde{i}})$ with the same identity; $d_{k,i,\tilde{i}}$ $(d'_{k,i,\tilde{i}})$ is the Euclidean distance between $\boldsymbol{f}_{k,i}$ $(\boldsymbol{f}'_{k,i})$ and the feature $\boldsymbol{f}_{k,j}$ $(\boldsymbol{f}'_{k,j})$ with different identities. The above $\tau_1$, $\tau_2$ and $\tau_3$ are margin hyperparameters. This paper empirically sets them to 0.3.

### *D. Single-Branch Dual-Stream Collaborative Network*

SDCN is mainly composed of an encoder $\boldsymbol{E}_2$, an attention layer, and a classifier $\boldsymbol{W}_2$. The network has strong robustness to occluded pedestrians by improving defense against the adversarial attacks of occluded samples. The network training is completed under the guidance of the verification guidance network. Specifically, the training data input to SDCN comes from $\boldsymbol{D}_t$ and $\boldsymbol{D}'_t$ respectively. The training samples from $\boldsymbol{D}_t$ are mainly used to make the network have better performance on unoccluded pedestrian images. The training samples from $\boldsymbol{D}'_t$ are used as adversarial examples to improve the model's robustness against obstacles. In addition, in SDCN training, only training samples from $\boldsymbol{D}_t$ are input to the pre-trained verification guidance network to assist SDCN in improving its ability to defend against adversarial attacks. If SDCN can effectively resist the interference caused by obstacles, then the features it extracted from the adversarial example $\boldsymbol{x}'_i$ should not deviate from $\boldsymbol{E}_1(\boldsymbol{x}_i)$ too much. Specifically, let $\boldsymbol{E}^l_1$ and $\boldsymbol{E}^l_2$ be the $l$-th layer of the encoders $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ respectively. To ensure SDCN can extract the features from the adversarial example $\boldsymbol{x}'_i$ roughly consistent with the full pedestrian image sample $\boldsymbol{x}_i$, the following consistency loss is used to optimize the parameters of SDCN as follows.

$$\boldsymbol{L}^l_{cl}(\boldsymbol{E}_2) = \|\boldsymbol{E}^l_2(\boldsymbol{x}'_i) - \boldsymbol{E}^l_1(\boldsymbol{x}_i)\|^2_F, \tag{4}$$

where $l = 3, 4$. If the feature extracted by the encoder $\boldsymbol{E}_2$ from $\boldsymbol{x}'_i$ has a large deviation from the feature extracted from $\boldsymbol{x}_i$, it indicates that $\boldsymbol{E}_2$ pays too much attention to the occluded area, which is not conducive to pedestrian identity matching. This problem can be avoided by using the $l_2$-loss between the occluded and unoccluded pedestrian image pairs. Since the feature $\boldsymbol{E}^l_1(\boldsymbol{x}_i)$ of the full pedestrian image sample $\boldsymbol{x}_i$ is fixed, $\boldsymbol{E}_2$ has certain ability to predict and restore the features of the occluded area under the constraint of the $l_2$-loss.

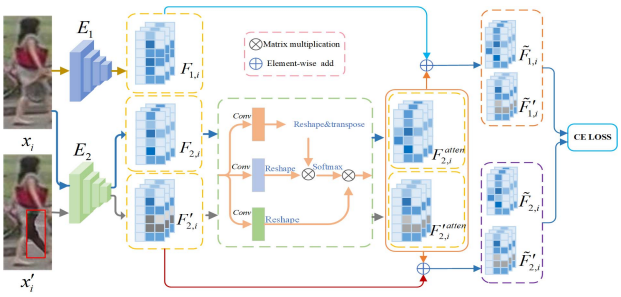



Fig. 4. Attention-based feature verification mechanism.

In addition, the $l_2$-loss allows for small deviations between $\boldsymbol{E}^l_2(\boldsymbol{x}'_i)$ and $\boldsymbol{E}^l_1(\boldsymbol{x}_i)$, which may introduce some information that is not beneficial for the re-ID task. Therefore, it is necessary to highlight the role of more critical information and reduce the attention to other information. The self-attention mechanism [33], [34] can represent the current feature by assigning a weight to each feature vector and linearly combining these feature vectors with these weights. The weight value of each feature vector is usually the inner product of the feature vector and the current feature vector. If a feature vector involved in the calculation is critical to the current task, the weight value of the vector is relatively large, and vice versa. Therefore, self-attention can focus on information that is more critical to the current task, reduce attention to other information, and even filter out irrelevant information.

To further suppress task-independent information introduced by obstacles or the $l_2$-loss in Eq. (4) and verify SDCN's ability to defend against adversarial attacks, an attention-based feature verification mechanism is designed (as shown in Fig.4). Suppose the features obtained after passing the samples $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$ through the encoders $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are $\boldsymbol{F}_{1,i} = \boldsymbol{E}_1(\boldsymbol{x}_i)$, $\boldsymbol{F}_{2,i} = \boldsymbol{E}_2(\boldsymbol{x}_i)$, and $\boldsymbol{F}'_{2,i} = \boldsymbol{E}_2(\boldsymbol{x}'_i)$, respectively. $\boldsymbol{F}_{2,i}$, and $\boldsymbol{F}'_{2,i}$ are input into the self-attention layer to obtain $\boldsymbol{F}^{atten}_{2,i}$ and $\boldsymbol{F}^{'atten}_{2,i}$, which are used to strengthen the role of useful information contained in $\boldsymbol{F}_{2,i}$, and $\boldsymbol{F}'_{2,i}$. $\boldsymbol{F}^{atten}_{2,i}$ is added as non-adversarial features to $\boldsymbol{F}_{1,i}$, so $\tilde{\boldsymbol{F}}_{1,i} = \boldsymbol{F}_{1,i} + \boldsymbol{F}^{atten}_{2,i}$ is obtained. Additionally, $\boldsymbol{F}^{'atten}_{2,i}$ as adversarial information is added to $\boldsymbol{F}_{1,i}$, so $\tilde{\boldsymbol{F}}'_{1,i} = \boldsymbol{F}_{1,i} + \boldsymbol{F}^{'atten}_{2,i}$ is obtained. For the input unoccluded pedestrian image $\boldsymbol{x}_i$, if both SDCN and verification guidance network can extract the consistent features, then $\tilde{\boldsymbol{F}}_{1,i}$ should be classified into the correct pedestrian identity through the classifier $\boldsymbol{W}_1$. In addition, if SDCN can defend the attacks from obstacles, the features extracted by SDCN from $\boldsymbol{x}'_i$ should be consistent with the features extracted from $\boldsymbol{x}_i$. In this case, the result obtained after the synthesis of $\boldsymbol{F}_{1,i}$ and $\boldsymbol{F}^{'atten}_{2,i}$ (i.e. $\boldsymbol{F}_{1,i} + \boldsymbol{F}^{'atten}_{2,i}$) should be still assigned to the correct identity when inputting to the pre-trained classifier $\boldsymbol{W}_1$. So, this paper uses the following cross-entropy-based verification loss to optimize the

parameters of SDCN:

$$L_{ce2}(\boldsymbol{E}_2) = -\frac{1}{n_b}\Bigg(\sum_{i=1}^{n_b}\sum_{l=1}^{L_t}\boldsymbol{I}_{(l=y_i)}\log(\boldsymbol{P}_l(\boldsymbol{W}_1(GAP(\tilde{\boldsymbol{F}}_{1,i})))) + \boldsymbol{I}_{(l=y_i)}\log(\boldsymbol{P}_l(\boldsymbol{W}_1(GAP(\tilde{\boldsymbol{F}}'_{1,i}))))\Bigg). \quad (5)$$

In the loss function, the first cross-entropy loss is mainly used to ensure that when the input of both SDCN and verification guidance network is the same unoccluded pedestrian image, the features extracted by SDCN and the features extracted by verification guidance network are compatible. Therefore, the classification results are not affected after $\boldsymbol{F}_{1,i} + \boldsymbol{F}_{2,i}^{atten}$, which ensures that SDCN also has a strong ability to recognize full pedestrian images. The second term of the loss function is mainly used to improve the robustness of SDCN to obstacles. If SDCN does not have an ability to defend the attacks from obstacles, the features extracted by SDCN are affected by obstacles and show incompatibility with unoccluded features $\boldsymbol{F}_{1,i}$. These features are mixed with $\boldsymbol{F}_{1,i}$ and sent to the pre-trained classifier $\boldsymbol{W}_1$, which definitely affects the correctness of the classification results. Therefore, the loss function in Eq. 5 plays a role of backward correction in SDCN training. Moreover, only GAP is used here, which is different from GAP+GMP used in Eq. 1. When the classifier $\boldsymbol{W}_1$ has been trained with GAP+GMP, if the features output by GAP are still correctly classified as pedestrian identities by $\boldsymbol{W}_1$, the encoder $\boldsymbol{E}_2$ must be able to extract richer salient features to compensate for the absence of GMP. This is beneficial to improve the ability of $\boldsymbol{E}_2$ to extract salient features. Therefore, only GAP is used in Eq. 5.

In order to ensure the effectiveness of the above mechanism, the samples from the dataset $\boldsymbol{D}'_t$ used in verification guidance network training and the samples from the dataset $\boldsymbol{D}'_t$ used in SDCN training are randomly generated during the training process. It ensures that the occluded samples used by the training classifier $\boldsymbol{W}_1$ are not identical to the occluded samples in Eq. 5. In addition, the output features of the attention layer and the original features are first added to achieve information enhancement, and then the added features are used as the final appearance features for testing. The following cross-entropy loss and triplet loss are used to optimize the parameters of SDCN. The cross-entropy loss is expressed as follows:

$$L_{ce3}(\boldsymbol{E}_2, \boldsymbol{W}_2) = -\frac{1}{n_b}\Bigg(\sum_{i=1}^{n_b}\sum_{l=1}^{L_t}\boldsymbol{I}_{(l=y_i)}\log \boldsymbol{P}_l(\boldsymbol{W}_2(\boldsymbol{f}_{2,s,i}))\Bigg)$$
$$L_{ce4}(\boldsymbol{E}_2, \boldsymbol{W}_2) = -\frac{1}{n_b}\Bigg(\sum_{i=1}^{n_b}\sum_{l=1}^{L_t}\boldsymbol{I}_{(l=y_i)}\log \boldsymbol{P}_l(\boldsymbol{W}_2(\boldsymbol{f}'_{2,s,i}))\Bigg), \quad (6)$$

where

$$\begin{aligned}\boldsymbol{f}_{2,s,i} &= GAP(\tilde{\boldsymbol{F}}_{2,i}) + GMP(\tilde{\boldsymbol{F}}_{2,i})\\ \boldsymbol{f}'_{2,s,i} &= GAP(\tilde{\boldsymbol{F}}'_{2,i}) + GMP(\tilde{\boldsymbol{F}}'_{2,i})\\ \tilde{\boldsymbol{F}}_{2,i} &= \boldsymbol{F}_{2,i} + \boldsymbol{F}_{2,i}^{atten}\\ \tilde{\boldsymbol{F}}'_{2,i} &= \boldsymbol{F}_{2,i} + \boldsymbol{F}_{2,i}^{'atten}.\end{aligned} \quad (7)$$

Additionally, the triplet loss is expressed as follows:

$$\begin{aligned}\boldsymbol{L}_{tri2}(\boldsymbol{E}_2, \boldsymbol{W}_2) = \frac{1}{n_b}\Bigg(&\sum_{i=1}^{n_b}[\max_{\tilde{i}}\tilde{d}_{a,i,\tilde{i}} - \min_j \tilde{d}_{a,i,j} + \varepsilon_1]_+\\ &+ \sum_{i=1}^{n_b}[\max_{\tilde{i}}\tilde{d}_{m,i,\tilde{i}} - \min_j \tilde{d}_{m,i,j} + \varepsilon_2]_+\\ &+ \sum_{i=1}^{n_b}[\max_{\tilde{i}}\tilde{d}_{s,i,\tilde{i}} - \min_j \tilde{d}_{s,i,j} + \varepsilon_3]_+\Bigg), \quad (8)\end{aligned}$$

where $\tilde{d}_{k,i,\tilde{i}} = \hat{d}_{k,i,\tilde{i}}$, $\hat{d}'_{k,i,\tilde{i}}$ $(k = a, m, s)$; $\hat{d}_{k,i,\tilde{i}}$ $(\hat{d}'_{k,i,\tilde{i}})$ is the Euclidean distance between $\boldsymbol{f}_{2,k,i}$ $(\boldsymbol{f}'_{2,k,i})$ and $\boldsymbol{f}_{2,k,\tilde{i}}$ $(\boldsymbol{f}'_{2,k,\tilde{i}})$ with the same identity; $\hat{d}_{k,i,j}$ $(\hat{d}'_{k,i,j})$ is the Euclidean distance between $\boldsymbol{f}_{2,k,i}$ $(\boldsymbol{f}'_{2,k,i})$ and $\boldsymbol{f}_{2,k,j}$ $(\boldsymbol{f}'_{2,k,j})$ with different identities. The above $\varepsilon_1$, $\varepsilon_2$ and $\varepsilon_3$ are margin hyperparameters. This paper empirically sets them to 0.3. Here, $\boldsymbol{f}_{2,a,i} = GAP(\boldsymbol{F}_{2,i}^{atten})$, $\boldsymbol{f}_{2,m,i} = GMP(\boldsymbol{F}_{2,i}^{atten})$.

### E. Optimization and Algorithms

The proposed method is composed of a verification guidance network and SDCN. The verification guidance network mainly provides verification guidance for SDCN. The network is pre-trained using the following loss function:

$$\boldsymbol{L}_1(\boldsymbol{E}_1, \boldsymbol{W}_1) = L_{ce1}(\boldsymbol{E}_1, \boldsymbol{W}_1) + L_{tri1}(\boldsymbol{E}_1, \boldsymbol{W}_1), \quad (9)$$

The loss functions used in SDCN training include consistency loss $L_{cl}^l(\boldsymbol{E}_2)$, verification loss $L_{ce2}(\boldsymbol{E}_2)$, discrimination loss $L_{ce3}(\boldsymbol{E}_2, \boldsymbol{W}_2)$, $L_{ce4}(\boldsymbol{E}_2, \boldsymbol{W}_2)$ and $L_{tri2}(\boldsymbol{E}_2, \boldsymbol{W}_2)$. So, the entire loss can be expressed as follows:

$$\begin{aligned}\boldsymbol{L}_2(\boldsymbol{E}_2, \boldsymbol{W}_2) = L_{tri2}(\boldsymbol{E}_2, \boldsymbol{W}_2) + \sum_{l=3,4}\lambda_1 L_{cl}^l(\boldsymbol{E}_2) + \lambda_2 L_{ce2}(\boldsymbol{E}_2)\\ + \lambda_3 L_{ce3}(\boldsymbol{E}_2, \boldsymbol{W}_2) + \lambda_4 L_{ce4}(\boldsymbol{E}_2, \boldsymbol{W}_2), \quad (10)\end{aligned}$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are four hyperparameters, which control the role of the corresponding loss function respectively. The complete procedure of model optimization is summarized in **Algo. 1**.

## IV. Experiments

### A. Datasets and Evaluation Protocol

This paper uses six challenging datasets to test the effectiveness of the proposed method. These datasets include Occluded-DukeMTMC [11], Occluded-ReID [10], Partial-REID [24] and Partial-iLIDS [25], as well as Market-1501 [35] and DukeMTMC-reID [36]. Additionally, the performance of the proposed method is compared with the corresponding performance of state-of-the-art methods to verify its advantages, and the effectiveness of each part of the proposed method is verified by ablation experiments.

**Occluded-DukeMTMC** was derived from DukeMTMC-reID. It was specially constructed for occluded person re-ID. The samples in this dataset were collected by eight non-overlapping cameras. The training set contains 15,618 images of 702 pedestrians. The testing set contains 19,871 images of 519 different pedestrians, each of which contains obstacles.

TABLE I

DATASET DETAILS AND SETTINGS USED IN EXPERIMENTS. PED: THE NUMBER OF PEDESTRIANS, IMG: THE NUMBER OF IMAGES, CAM: THE NUMBER OF CAMERAS. "-" MEANS THAT THE CORRESPONDING DATA IS NOT AVAILABLE

| Datasets | Ped | Training | | Gallery (Testing) | | Probe (Testing) | | Cam |
|---|---|---|---|---|---|---|---|---|
| | | Ped | Img | Ped | Img | Ped | Img | |
| Occluded-ReID | 200 | – | – | 200 | 1000 | 200 | 1000 | – |
| Occluded-DukeMTMC | 1,221 | 702 | 15,618 | 519 | 17,661 | 519 | 2,210 | 8 |
| Partial-REID | 60 | – | – | 60 | 300 | 60 | 300 | 4 |
| Partial-iLIDS | 119 | – | – | 119 | 119 | 119 | 119 | 2 |
| Market1501 | 1,501 | 751 | 12,936 | 750 | 19,732 | 750 | 3,368 | 6 |
| DukeMTMC-reID | 1,404 | 702 | 16,522 | 702 | 17,661 | 702 | 2,228 | 8 |

**Algorithm 1** Occluded Person Re-Identification via Defending Against Attacks From Obstacles (OPR-DAAO)

**Input:** Training images and their labels $\boldsymbol{D}_t = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_t}$,the maximum of $Iteration_1$ and $Iteration_2$.
**Output:** The trained $\boldsymbol{E}_2$, $\boldsymbol{W}_2$.
**Step I:** Train the verification guidance network.
Sample a batch of labeled source data to $\boldsymbol{E}_1$.
Initialize $\boldsymbol{E}_1$, $\boldsymbol{W}_1$
**for** $iter$=1, $\cdots$, $Iteration_1$ **do**
  Generate occlused data $\boldsymbol{D}_t' = \{\boldsymbol{x}_i', \boldsymbol{y}_i\}_{i=1}^{N_t}$.
  Update $\boldsymbol{E}_1$ and $\boldsymbol{W}_1$ by minimizing the loss in Eq.(1) and in Eq.(3).
**end for**
**Step II:**Train the SDCN network.
Load the learned $\boldsymbol{E}_1$, $\boldsymbol{W}_1$.
Initialize $\boldsymbol{E}_2$, $\boldsymbol{W}_2$.
**for** $iter$=1, $\cdots$, $Iteration_2$ **do**
  Update $\boldsymbol{E}_2$, $\boldsymbol{W}_2$ by minimizing the loss in Eqs. (4),(5),(6) and (8).
**end for**

**Occluded-ReID** contains 2,000 images of 200 occluded pedestrians. All the samples were collected by a mobile camera. Each pedestrian identity has five full-body images and five severely occluded images. The experimental settings in this paper are same as existing methods [16], [17], [37]. In the experiments, Market-1501 is used as the training set and Occluded-ReID is used as the testing set.

**Partial-REID** contains 600 images of 60 pedestrians collected by six non-overlapping cameras. Each pedestrian has five full-body images and five images after cropping occluded areas. The collection of cropped partial images is used as Probe, and the collection of full-body images is used as gallery. Due to the small scale of this dataset, a general strategy is adopted in the experiments, i.e., Market-1501 and Partial-REID are used as the training set and the testing set, respectively.

**Partial-iLIDS** Partial-iLIDS was derived from iLIDS [38]. This dataset contains 238 images of 119 pedestrians. All the images were captured by two non-overlapping cameras. 119 images were captured from the same camera view. The obstacles were cropped from these 119 images. The remaining 119 image are full-body images. In the experiments, 119 pedestrian images after cropping obstacles are used as Probe, and the remaining 119 images are used as gallery. Same as existing methods, the experiments use Market-1501 as the training set and Partial-iLIDS as the testing set.

**Market-1501** contains 32,668 images of 1501 pedestrians. These pedestrian images were captured by six non-overlapping cameras. In this dataset, the training set contains 12,936 images of 751 pedestrians, and the testing set contains 19,732 images of 750 pedestrians.

**DukeMTMC-reID** contains 36,411 images of 1,404 pedestrians, which were captured by eight non-overlapping cameras. In this dataset, the training set contains 16,522 images of 702 pedestrians, and the testing set contains 19,889 images of the remaining 702 pedestrians. The details of each dataset are shown in Tab. I.

**Evaluation Protocols**: This paper uses cumulative matching characteristic (CMC) [39] and mean average precision (mAP) [35] as the objective evaluation indicators to evaluate re-ID performance.

### *B. Implementation Details*

The proposed method uses ResNet50 [40] pre-trained on ImageNet [41] as the backbone of the two networks. Before inputting to the encoders, the size of all images is uniformly adjusted to $256 \times 128$, and the ADM optimizer [42] is used to update the network parameters. Additionally, similar to the method in [43], random flipping and color dithering are used to achieve data augmentation. In the experiments, the batch size $n_b$ is set to 32, each batch contains eight pedestrians, and each pedestrian has four samples. The initial learning rates of the encoders $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ and the classifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are set to 0.0002. The training of both verification guidance network and SDCN requires 150 epochs. In $0 \sim 10$ epochs, the warm-up strategy in [44] is used to adjust the learning rate. Starting from the 11-th epoch, the learning rate remains the same until the 40-th epoch. At the 41-th epoch, the learning rate decays by 10%, and then remains unchanged until the 70-th epoch. At the 71-th epoch, the learning rate decays again by 10% and remains unchanged until the 150-th epoch. In the experiments, the hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are set to 0.01, 0.1, 0.1 and 1.0, respectively. The impact of hyperparameters on performance will be specified in Section 4.5. The proposed model was implemented under the pytorch framework [45]. All experiments were done on a NVIDIA GeForce RTX2080Ti GPU platform.

TABLE II
ABLATION STUDY OF EACH COMPONENT OF THE PROPOSED METHOD. BOTH CMC AND MAP RATE (%) OBTAINED BY EACH METHOD ARE LISTED

| Methods | Rank-1 | Rank-5 | Rank-10 | mAP |
|---|---|---|---|---|
| Baseline | 49.0 | 64.3 | 69.8 | 42.6 |
| Baseline+ASS | 53.8 | 72.5 | 78.7 | 44.5 |
| Baseline+ASS+SAT | 55.5 | 73.7 | 79.7 | 45.9 |
| Baseline+ASS+CL | 63.1 | 75.6 | 82.1 | 47.3 |
| Baseline+ASS+CL+VL | 65.1 | 77.6 | 83.2 | 48.6 |
| Baseline+ASS+CL(Be)+SAT | 64.4 | 76.2 | 82.8 | 48.0 |
| Baseline+ASS+CL(Be)+SAT+VL | 66.2 | 78.4 | 83.9 | 55.4 |
| Baseline+ASS+CL(Af)+SAT+VL | 64.8 | 76.7 | 82.3 | 47.5 |

### C. Ablation Study

The proposed OPR-DAAO is mainly composed of adversarial sample synthesis, verification guidance network, and SDCN. SDCN includes three parts, consistency loss (CL) used to constrain integrity feature extraction, self-attention (SAT), and verification loss (VL) used to defend adversarial attacks. In this paper, the ResNet50 trained on the full pedestrian images under the constraint of the loss function $\boldsymbol{L}_1(\boldsymbol{E}_1, \boldsymbol{W}_1) = L_{ce1}(\boldsymbol{E}_1, \boldsymbol{W}_1) + L_{tri1}(\boldsymbol{E}_1, \boldsymbol{W}_1)$ is used as "Baseline". Under the constraint of the same loss function, the ResNet50 trained on the full pedestrian images and the adversarial examples synthesized (ASS) by the proposed method is termed as "Baseline+ASS". Moreover, in Tab.II, "Baseline+ASS+CL" means that the consistency loss is added to "Baseline+ASS" (Eq. 5). "Baseline+ASS+SAT" means that the self-attention is added to "Baseline+ASS", that is, the verification guidance network does not exist. "Baseline+ASS+CL+VL" indicates that VL is added to "Baseline+ASS+CL". "Baseline+ASS+CL(Be)+SAT" means that the self-attention layer is added to "Baseline+ASS+CL" and placed after "CL". "Baseline+ASS+CL(Be)+SAT+VL" means that the verification loss is added to "Baseline+ASS+CL(Be)+SAT". "Baseline+ASS+CL(Af)+SAT" means CL is placed after SAT. The above methods were all trained on Occluded-DukeMTMC under the constraints of the loss functions shown in Eqs. (6) and (8). The experimental results are shown in Tab.II.

*1) The Effectiveness of Adversarial Sample Synthesis:* According to the re-ID performance obtained by Baseline and Baseline+ASS shown in Tab.II, the performance of the proposed model is effectively improved by using the generated adversarial samples to participate in the training of the feature extraction network. Specifically, after adding generated samples, the re-ID rate corresponding to Rank-1 increases from 49.0% to 53.8%, and the accuracy rate corresponding to mAP increases from 42.6% to 44.5%. The above results confirm that adversarial samples have a certain ability to prompt the feature extraction network to extract robust features.

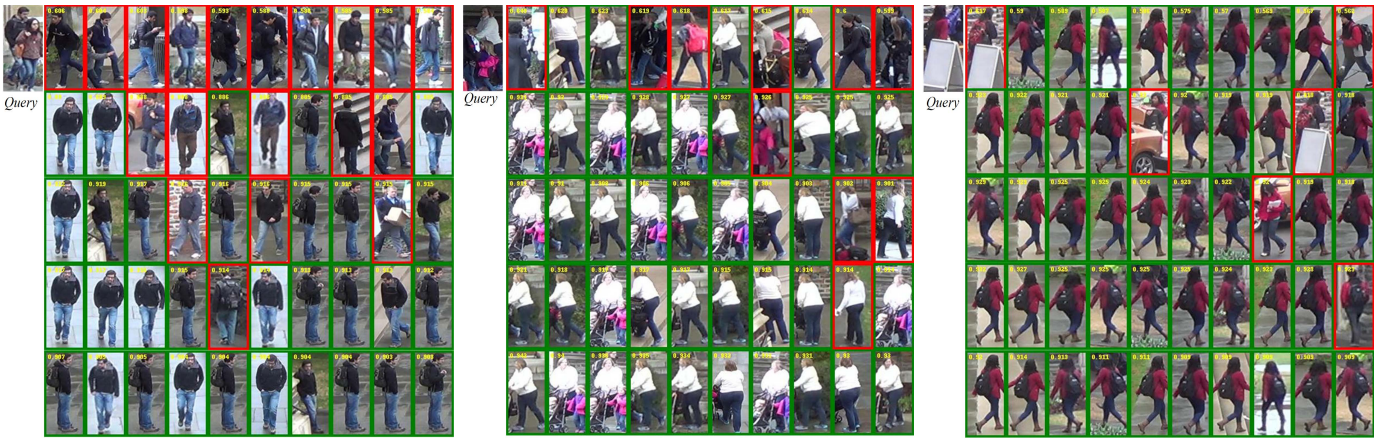
*2) The Effectiveness of Consistency Loss:* Consistency loss is mainly used to enable SDCN to extract features from occluded pedestrian images that are roughly consistent with unoccluded pedestrian images. According to the experimental results shown in Tab. II, after the CL is added to Baseline+ASS model, the re-ID rate corresponding to Rank-1 increases from 53.8% to 63.1%, and the accuracy rate corresponding to mAP increases from 44.5% to 47.3%. The increase in Rank-1 and mAP reaches 9.7% and 2.8% respectively, which confirms that consistency loss can effectively improve the performance of SDCN with the assistance of the verification guidance network. According to the performance comparison of "Baseline+ASS+CL(Be)+SAT+VL" and "Baseline+ASS+CL(Be)+SAT+VL", it is reasonable to place SAT after CL.

*3) The Effectiveness of SAT:* The SAT layer is introduced to strengthen the role of useful information. According to the experimental results shown in Tab. II, after adding the SAT module in "Baseline+ASS" ("Baseline+ASS+CL"), the re-ID rate on Rank-1 increases from 53.8(63.1)% to 55.5(63.9)% and the accuracy rate on mAP increases from 44.5(47.3)% to 45.9(48.0)%. Therefore, the above-mentioned results confirm that the SAT layer can make SDCN pay more attention to effective pedestrian features.

*4) The Effectiveness of Verification Loss:* The verification loss is introduced to use the dual-stream cooperative network to make SDCN have a strong ability to defend the attacks from obstacles. According to the experimental results shown in Tab. II, after introducing the verification loss on the basis of Baseline+ASS+CL(Be)+SAT, the re-ID performance of the model on Rank-1 increases from 64.4% to 66.2%, and the accuracy rate on mAP increases from 48.0% to 55.4%. The corresponding increase rate reaches 1.8% and 7.4%, respectively. This confirms that the dual-stream cooperative network can guide SDCN to defend against the attacks from obstacles. According to the performance comparison of "Baseline+ASS+SAT " and " Baseline+ASS+CL(Be)+SAT+VL", the verification guidance network (CL(Be)+VL) plays a positive role in improving performance. Additionally, Fig.5 shows the visualization of the retrieval results under different functional modules. According to these results, conclusions consistent with the above analysis can be drawn.

### D. Comparison With State-of-the-Art Methods

*1) Experiments on Occluded-DukeMTMC:* In order to verify the performance of the proposed method, the proposed method is first applied to the commonly used occluded dataset Occluded-Duke, and its performance is compared with the corresponding performance of state-of-the-art methods. The comparative methods include PartBili [67], FDGAN [47], PartAlign [46], PCB [48], DSR [25], PGFA [11], AdOccl [49], SORN [18], MHSANet [37], CBDBNet [50], HOReID [17], RFCnet [51], PGFL-KD [20], Pirt [52], and PAT [23]. According to the experimental results shown in Tab.III, since the first five methods do not consider the impact of obstacles on re-ID performance, the corresponding re-ID rate on Rank-1 and mAP is low. AdOccl, PGFA, HOReID, RFCnet, PGFL-KD, Pirt, PAT and MHSANet all consider the impact of obstacles on re-ID performance. Therefore, compared with the




(a) (b) (c)

Fig. 5. Retrieval results under different functional modules. For each sub-image, the first line shows the retrieval results under Baseline, the second line shows the retrieval results of Baseline+ASS, the third line shows the retrieval results of Baseline+ASS+CL, the fourth line shows the retrieval results of Baseline+ASS+CL+SAT, and the fifth line shows the retrieval results of Baseline+ASS+CL+SAT+VL. The green box means the matching result is correct, and the red box means the matching result is wrong.

TABLE III

COMPARISON OF EXPERIMENTAL RESULTS ON OCCLUDED-DUKEMTMC BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART OCCLUDED PERSON RE-ID METHODS. THE CMC AND MAP RATES (%) OF EACH METHOD ARE REPORTED. '–' MEANS NO RESULTS ARE REPORTED. THE BEST RESULTS ARE MARKED IN BOLD, AND THE SECOND-BEST RESULTS ARE MARKED IN BLUE

| Methods | Rank-1 | Rank-5 | Rank-10 | mAP |
|---|---|---|---|---|
| PartBili [67] | 36.9 | – | – | – |
| FDGAN [47] | 40.8 | – | – | – |
| PartAlign [46] | 28.8 | 44.6 | 51.0 | 20.2 |
| PCB [48] | 42.6 | 57.1 | 62.9 | 33.7 |
| DSR [25] | 40.8 | 58.2 | 65.2 | 30.4 |
| PGFA [11] | 51.4 | 68.6 | 74.9 | 37.3 |
| AdOccl [49] | 44.5 | – | – | 32.2 |
| SORN [18] | 57.6 | 73.3 | 79.0 | 46.3 |
| MHSANet [37] | 55.4 | 70.2 | 76.4 | 42.4 |
| CBDBNet [50] | 50.9 | 66.0 | 66.0 | 38.9 |
| HOReID [17] | 55.1 | – | – | 43.8 |
| RFCnet [51] | 63.9 | 77.6 | 82.1 | 54.5 |
| PGFL-KD [20] | 63.0 | – | – | 54.1 |
| Pirt [52] | 60.0 | – | – | 50.9 |
| PAT [23] | 64.5 | – | – | 53.6 |
| **OPR-DAAO** | **66.2** | **78.4** | **83.9** | **55.4** |

TABLE IV

COMPARISON OF EXPERIMENTAL RESULTS ON OCCLUDED-REID BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART PARTIAL PERSON RE-ID METHODS. THE CMC AND MAP RATES (%) OF EACH METHOD ARE REPORTED. '–' MEANS NO RESULTS ARE REPORTED. THE BEST RESULTS ARE MARKED IN BOLD, AND THE SECOND-BEST RESULTS ARE MARKED IN BLUE

| Methods | Rank-1 | Rank-5 | mAP |
|---|---|---|---|
| PCB [48] | 59.3 | 75.2 | 53.2 |
| DSR [25] | 72.8 | – | 62.8 |
| OPR [10] | 68.1 | – | – |
| OSNet [53] | 39.7 | 57.9 | 36.0 |
| FPR [22] | 78.3 | – | 68.0 |
| GASM [54] | 74.5 | – | 65.6 |
| HOReID [17] | 80.3 | – | 70.2 |
| PVPM [16] | 70.4 | – | 65.6 |
| PGFL-KD [20] | 80.7 | – | 70.3 |
| PAT [23] | 81.6 | – | 72.1 |
| **OPR-DAAO** | **84.2** | **87.3** | **75.1** |

first five methods, the re-ID accuracy on Rank-1 and mAP of these methods are significantly improved. Among these methods, the latest PAT achieves the best re-ID performance. Its re-ID rate on Rank-1 and mAP reaches 64.5% and 53.6%, respectively. Compared with PAT, the re-ID rate of the proposed method on Rank-1 and mAP is further improved, which reaches 66.2% and 55.4% respectively. This not only proves the effectiveness of the proposed method, but also verifies the superiority of the proposed method over all comparative methods.

*2) Experiments on Occluded-ReID:* In order to further verify the effectiveness of the proposed method on occluded person re-ID, Occluded-ReID is used to test its performance, and its performance is compared with state-of-the-art methods. The experimental results are shown in Tab.IV. In this experiment, the comparative methods include PCB [48], DSR [25], OPR [10], OSNet [53], FPR [22], GASM [54], HOReID [17], PVPM [16], PGFL-KD [20], and PAT [23]. According to the experimental results shown in Tab.IV, the re-ID accuracy of Rank-1 and mAP obtained by the proposed method on this dataset reaches 84.2% and 75.1%, respectively. The re-ID accuracy of Rank-1 and mAP obtained by the second-best method PAT reaches 81.6% and 72.1%, respectively, which is lower than the corresponding performance obtained by the proposed method. This proves that the proposed method has better re-ID performance on occluded datasets. Additionally, it is effective to introduce adversarial attacks into the design of occluded person re-ID models.

TABLE V

COMPARISON OF EXPERIMENTAL RESULTS ON PARTIAL-REID AND PARTIAL-iLIDS BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART PERSON RE-ID METHODS. THE CMC AND MAP RATES (%) OF EACH METHOD ARE REPORTED. "-" MEANS NO RESULTS ARE REPORTED. THE BEST RESULTS ARE MARKED IN BOLD, AND THE SECOND-BEST RESULTS ARE MARKED IN BLUE

| Methods | Partial-REID | | | Partial-iLIDS | | |
|---|---|---|---|---|---|---|
| | Rank-1 | Rank-3 | mAP | Rank-1 | Rank-3 | mAP |
| DSR [25] | 50.7 | 70.0 | – | 58.8 | 67.2 | – |
| HACNN [55] | 37.0 | – | 40.4 | – | – | – |
| MLFN [56] | 42.7 | – | 45.7 | – | – | – |
| PCB [48] | 56.3 | – | 54.7 | 46.8 | – | 40.2 |
| OPR [10] | 78.5 | – | – | – | – | – |
| VPM [44] | 67.7 | 81.9 | – | 65.5 | 74.8 | – |
| FPR [22] | 81.0 | – | 76.6 | 68.1 | – | 61.8 |
| OSNet [53] | 48.7 | – | 49.3 | – | – | – |
| PGFA [11] | 68.0 | 80.0 | – | 69.1 | 80.9 | – |
| MHSANet [37] | 81.3 | 87.7 | – | 73.6 | 85.4 | – |
| PVPM [16] | 78.3 | – | – | – | – | – |
| CBDBNet [50] | 66.7 | 78.3 | – | 68.4 | 81.5 | – |
| HOReID [17] | 85.3 | 91.0 | – | 72.6 | 86.4 | – |
| STNReID [27] | 66.7 | 80.3 | – | 54.6 | 71.3 | – |
| PGFL-KD [20] | 85.1 | 90.8 | – | 74.0 | **86.7** | – |
| **OPR-DAAO** | **86.5** | **91.3** | **80.1** | **76.2** | 85.1 | **69.1** |

*3) Experiments on Partial-REID and Partial-iLIDS:* Although the proposed method is designed for occluded person re-ID, it is also effective for partial person re-ID. This paper uses two datasets Partial-REID and Partial-iLIDS to test the performance of the proposed method. In this experiment, the proposed method is compared with DSR [25], HACNN [55], MLFN [56], PCB [48], OPR [10], VPM [44], FPR [22], OSNet [53], PGFA [11], MHSANet [37], PVPM [16], CBDBNet [50], HOReID [17], STNReID [27] and PGFL-KD [20]. Tab.V shows the re-ID accuracy of different methods on Rank-1 and mAP. The proposed method can also obtain excellent re-ID performance on Partial person re-ID. Compared with the second-best method PGFL-KD, the re-ID accuracy of Rank-1 and mAP obtained by the proposed method reaches 86.5% and 80.1% (76.2% and 69.1%) respectively on Partial-REID (Partial-iLIDS), which exceeds the re-ID accuracy obtained by the latest method PGFL-KD. This confirms that the proposed method also achieves excellent re-ID performance on partial data.

*4) Experiments on Market-1501 and DukeMTMC-reID:* Although the proposed method is specifically designed for occluded person re-ID, it can be still applied to full person re-ID datasets. So, this paper uses two full pedestrian datasets, Market-1501 and DukeMTMC-reID, to test the performance of the proposed method. In this experiment, the proposed method is compared with MLFN [56], MCAM [57], Mancs [58], HACNN [55], FANN [59], PAN [60], IANet [61], PGFA [11], SORN [18], HOReID [17], HACNDHA [62], MHSANet [37], CSPRNet [63], PEFB [64], JAD [65] and AOPS+VRM [66]. As shown in Tab.VI, for Market-1501 and DukeMTMC-reID, the re-ID rates on Rank-1 obtained by the proposed method (OPR-DAAO) are 95.1% and 88.5% respectively, and the accuracy rates on mAP obtained by OPR-DAAO are 86.2% and 76.7% respectively. Compared with the performance of the latest methods PEFB and AOPS+VRM, the proposed method has stronger competitiveness. This further confirms the effectiveness of the proposed method.

TABLE VI

COMPARISON OF EXPERIMENTAL RESULTS ON MARKET-1501 AND DUKEMTMC-reID BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS. THE CMC AND MAP RATES (%) OF EACH METHOD ARE REPORTED. "-" MEANS NO RESULTS ARE REPORTED. THE BEST RESULTS ARE MARKED IN BOLD, AND THE SECOND-BEST RESULTS ARE MARKED IN BLUE

| Methods | Market-1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| MLFN [56] | 90.0 | 74.3 | 81.0 | 62.8 |
| MCAM [57] | 83.8 | 74.3 | – | – |
| Mancs [58] | 93.1 | 82.3 | 84.9 | 71.8 |
| HACNN [55] | 91.2 | 75.7 | 80.5 | 63.8 |
| FANN [59] | 90.3 | 76.1 | – | – |
| PAN [60] | 82.8 | 63.4 | 71.6 | 51.5 |
| IANet [61] | 94.4 | 83.1 | 87.1 | 73.4 |
| PGFA [11] | 91.2 | 76.8 | 82.6 | 65.5 |
| SORN [18] | 94.8 | 84.5 | 86.9 | 74.1 |
| HOReID [17] | 94.2 | 84.9 | 86.9 | 75.6 |
| HACNDHA [62] | 91.3 | 76.0 | 81.3 | 64.1 |
| MHSANet [37] | 94.6 | 84.0 | 87.3 | 73.1 |
| CSPRNet [63] | 94.2 | 84.8 | 83.5 | 71.9 |
| PEFB [64] | 92.7 | 81.3 | 86.2 | 72.6 |
| JAD [65] | 88.7 | 70.3 | 77.2 | 57.8 |
| AOPS+VRM [66] | 94.6 | 85.3 | 87.5 | 76.3 |
| **OPR-DAAO** | **95.1** | **86.2** | **88.5** | **76.5** |

TABLE VII

COMPARISON OF EXPERIMENTAL RESULTS ON OCCLUDED-DUKE BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART OCCLUDED PERSON RE-ID METHODS. THE CMC AND MAP RATES (%) OF EACH METHOD ARE REPORTED. "–" MEANS NO RESULTS ARE REPORTED. THE BEST RESULTS ARE MARKED IN BOLD, AND THE DATA IN PARENTHESIS IS THE PERFORMANCE OF THE CORRESPONDING MODEL TRAINED ON THE ORIGINAL TRAINING SET

| Methods | Rank-1 | Rank-5 | mAP |
|---|---|---|---|
| DSR [25] | 56.2(52.1) | 71.0(67.8) | 47.6(44.5) |
| PGFA [11] | 53.1(51.4) | 70.2(68.6) | 39.2(37.3) |
| PVPM [16] | 52.3(50.6) | 65.1(62.1) | 35.6(33.4) |
| HOReID [17] | 56.7(55.1) | 67.5(–) | 45.0(43.8) |
| **OPR-DAAO** | **66.2** | **78.4** | **55.4** |

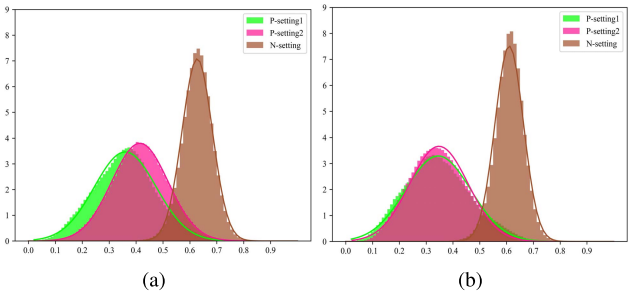



Fig. 6. Changes in the performance of the proposed method under different training and testing settings. Among them, the horizontal axis represents the distance between features of two images in the gallery, and the vertical axis represents the number of samples that fall on the corresponding distance.

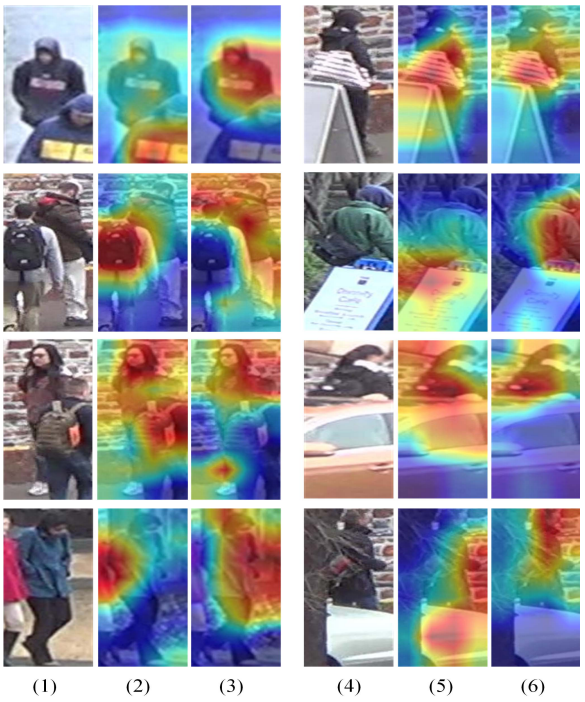



Fig. 7. The effectiveness of adversarial attack and defense. The first and fourth columns show the occluded pedestrian images. The second and fifth columns show the focused areas of the proposed model trained on the original training set of Occluded-DukeMTMC. The third and sixth columns show the focused areas of the proposed model trained on the adversarial samples and the original training samples of Occluded-DukeMTMC. The warmer color means the corresponding area received more attention of the proposed model.

### *E. Further Discussion*

In the above experiments, the proposed method includes the occluded sample synthesis. This step provides richer training samples for the training set. To verify the impact of the adversarial samples generated by the proposed method on the performance of existing methods, the training data generated by the proposed method and the original training data are applied to the modeling training of some existing methods. As the occluded person re-ID methods, the codes of DSR [25], PGFA [11], PVPM [16] and HOReID [17] are public, so the corresponding experimental results were obtained smoothly in this process. Therefore, the performance of the proposed method is only compared with DSR, PGFA, PVPM and HOReID in this experiment. According to the experimental results shown in Tab.VII, the adversarial examples mixed with the original training samples can effectively improve the re-ID performance of all methods due to data augmentation. Nevertheless, the proposed method is still significantly superior to the comparative methods. This is mainly because the designed verification guidance adversarial training framework effectively utilizes the guidance and supervision of full person images. The excellent performance confirms that the proposed method is more robust than existing methods.

Figs. 6 and 7 further demonstrate the effectiveness of the proposed method. Fig. 6 shows the distance distribution of pedestrian image features from different cameras in the gallery set when the proposed method is trained on different training sets. In P-setting1 of Fig. 6(a), both training set and gallery set are from Occluded-DukeMTMC, and the distribution map along horizontal axis is the distance between the features of positive samples from different camera views in the gallery set. In P-setting2 of Fig.6(a), the training set is the same as that used in P-setting1, but the testing samples are occlusion samples synthesized on the gallery set of P-setting1, and the distribution map along horizontal axis is the distance between features of positive samples of synthesized occlusion images from different camera views in the gallery set. N-setting presents the distance distribution of negative pair features in the gallery of Occluded-DukeMTMC. Under N-setting, the model is trained on the training set of Occluded-DukeMTMC. If the model is resistant to occlusion attacks, the distributions of P-setting1 and P-setting2 should be coincident. Since the adversarial examples do not participate in model training, the model does not show a strong ability to resist occlusion attacks.

In Fig.6(b), P-setting1 indicates the model is trained on original and synthesized training set. The samples involved in the calculation of the distribution map are the same as those in P-setting1 of Fig. 6(a). In P-setting2 of Fig.6(b), the training set is the same as that used in P-setting1 of Fig.6(b), but the testing set is the same as that used in P-setting2 of Fig.6(a). In N-setting, the model is trained on the training set of P-setting1 of Fig.6(b), and its distance distribution is obtained

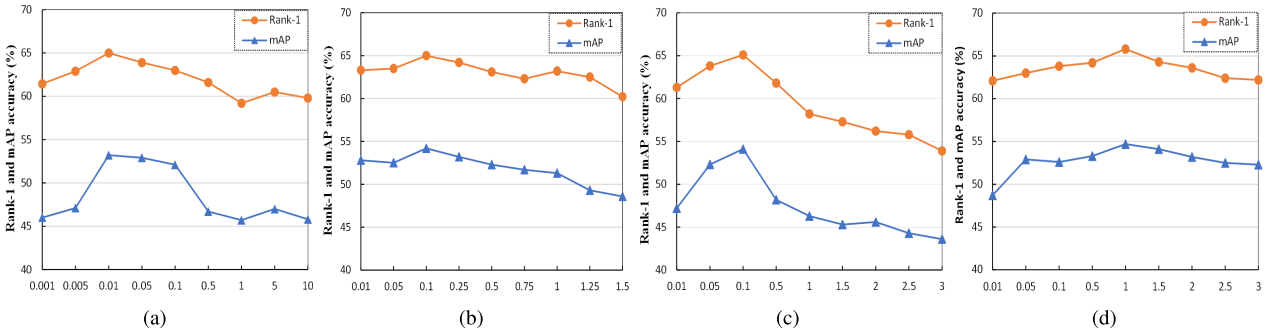



Fig. 8. Impact analysis of hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ on model performance. (a) The impact of $\lambda_1$ on model performance, (b) The impact of $\lambda_2$ on model performance, (c) The impact of $\lambda_3$ on model performance, (d) The impact of $\lambda_4$ on model performance.

by calculating the distance between negative sample features of the gallery set of Occluded-DukeMTMC. If the model is able to resist occlusion interferences, the distribution maps of P-setting1 and P-setting2 should be consistent. As shown in Fig. 6(b), the distribution maps of the proposed method on occluded datasets is consistent with that on non-occluded datasets after using the synthesized occlusion samples and the original samples to train the proposed model. Additionally, the intersection area between the red area and the brown area in Fig. 6(b) is much smaller than that in Fig. 6(a), which confirms that the proposed method can effectively reduce the number of hard samples. Besides, the results in Fig. 7 show that after training the model with the synthesized adversarial examples and the original samples, the obtained model is more effective than the model trained only using original samples. This indicates that it is effective to use the idea of adversarial attack and defense to solve the problem of occluded person re-ID.

### *F. Parameter Analysis*

The proposed method contains four hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$. In this section, the impact of $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ on model performance is analyzed on Occluded-Duke.

*1) The Impact of $\lambda_1$:* In Eq. 10, the consistency loss $L_{cl}^{l}(\boldsymbol{E}_2)$ is used to ensure that $\boldsymbol{E}_2(\boldsymbol{x}_i')$ does not deviate from $\boldsymbol{E}_1(\boldsymbol{x}_i)$ too much, but it cannot ensure that $\boldsymbol{E}_2(\boldsymbol{x}_i') = \boldsymbol{E}_1(\boldsymbol{x}_i)$. Differences between $\boldsymbol{E}_2(\boldsymbol{x}_i')$ and $\boldsymbol{E}_1(\boldsymbol{x}_i)$ are allowed. $\boldsymbol{E}_2(\boldsymbol{x}_i')$ may have a certain estimation error, so $\lambda_1$ cannot be set too large. Fig.8(a) shows the change in model performance when $\lambda_1$ increases from 0.001 to 10. When $\lambda_1 \in [0.001, 0.05]$, the proposed method shows good performance; when $\lambda_1 = 0.01$, the proposed method achieves the highest re-ID rate on Rank-1 and the highest accuracy rate on mAP, and when $\lambda_2$ is greater than 0.01, the model performance slightly decreases. Therefore, this paper sets $\lambda_1 = 0.01$ in all experiments.

*2) The Impacts of $\lambda_2$ and $\lambda_3$:* In Eq. 10, the cross entropy losses $L_{ce2}(\boldsymbol{E}_2)$ and $L_{ce3}(\boldsymbol{E}_2, \boldsymbol{W}_2)$ use the samples from $\boldsymbol{D}_t \bigcup \boldsymbol{D}_t'$ and $\boldsymbol{D}_t$ respectively to train the encoder $\boldsymbol{E}_2$, so that $\boldsymbol{E}_2$ has a strong ability to extract discriminative features. Within a certain range, the larger values of $\lambda_2$ and $\lambda_3$ mean the greater ability of $\boldsymbol{E}_2$ to extract discriminative features. Fig.8(b) shows the change in model performance when $\lambda_2$ takes different values. As shown in Fig.8(b), when $\lambda_2$ increases from 0.01 to 0.1, the model performance gradually improves. But when is greater than 0.1, the model performance decreases. Therefore, this paper sets $\lambda_2 = 0.1$ in all experiments. Fig.8(c) shows the change in model performance when $\lambda_3$ increases from 0.01 to 3. When $\lambda_3 \in [0.05, 0.1]$, the proposed method performs good performance, and when $\lambda_3 = 0.1$, the proposed method obtains the highest recognition rate on both Rank-1 and mAP, while when $\lambda_3$ is greater than 0.1, the model performance decreases. Therefore, this paper sets $\lambda_3 = 0.1$ in all experiments.

*3) The Impact of $\lambda_4$:* In Eq. (10), $\lambda_4$ is used to adjust the role of the discrimination loss $L_{ce4}(\boldsymbol{E}_2, \boldsymbol{W}_2)$. In $L_{ce4}(\boldsymbol{E}_2, \boldsymbol{W}_2)$, only the samples in $\boldsymbol{D}_t'$ are used to train $\boldsymbol{E}_2$. To make $\boldsymbol{E}_2$ perform well on $\boldsymbol{x}_i' \in \boldsymbol{D}_t'$, $\lambda_4$ should be larger than the values of other hyperparameters. Fig.8(d) shows the impact of $\lambda_4$ on model performance when $\lambda_4$ increases from 0.01 to 3. When $\lambda_4$ changes from 0.1 to 1, the model performance gradually improves. When $\lambda_4$ reaches 1.0, the corresponding performance reaches a peak, so this paper sets $\lambda_4$ to 1.0 in the experiments.

## V. Conclusion

This paper redesigns one effective occluded pedestrian re-ID method from a brand-new view. The proposed method is mainly composed of adversarial example synthesis, verification guidance network, and SDCN. Under the guidance and cooperation of the pre-trained verification guidance network, SDCN training is realized by using full pedestrian samples and synthesized occluded samples, thereby giving SDCN the ability to defend against the attacks from adversarial examples (occluded pedestrian samples). Unlike existing methods, the proposed method is no longer dedicated to extracting features from unoccluded human body areas to solve the issues of occluded person re-ID. Instead, it treats occluded pedestrian samples as adversarial example and uses them to implement adversarial training of the re-ID model. If the model can defend the attacks of adversarial examples, its generalization ability on occluded pedestrian images is significantly improved. Additionally, there is no need to use any external models (such as key point detection model, human body analysis model) to assist re-ID model training, so it can effectively avoid the negative impact of external model detection results on re-ID model training. The experimental results prove the effectiveness of the

proposed method and its superiority over state-of-the-art methods.

## REFERENCES


[1] W. Ding, X. Wei, R. Ji, X. Hong, Q. Tian, and Y. Gong, "Beyond universal person re-identification attack," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3442–3455, 2021.

[2] 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 Security*, vol. 16, pp. 1480–1494, 2021.

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

[4] Z. Zhu et al., "Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification," *J. Vis. Commun. Image Represent.*, vol. 80, May 2021, Art. no. 103303.

[5] D. Wu, M. Ye, G. Lin, X. Gao, and J. Shen, "Person re-identification by context-aware part attention and multi-head collaborative learning," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 115–126, 2022.

[6] H. Li, J. Pang, D. Tao, and Z. Yu, "Cross adversarial consistency self-prediction learning for unsupervised domain adaptation person re-identification," *Inf. Sci.*, vol. 559, pp. 46–60, Jun. 2021.

[7] M. Ye, C. Chen, J. Shen, and L. Shao, "Dynamic tri-level relation mining with attentive graph for visible infrared re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 386–398, 2022, doi: 10.1109/TIFS.2021.3139224.

[8] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022, doi: 10.1109/TPAMI.2021.3054775.

[9] H. Li, N. Dong, Z. Yu, D. Tao, and G. Qi, "Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2814–2830, May 2022.

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

[11] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 542–551.

[12] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.

[13] H. Huang et al., "EaNet: Enhancing alignment for cross-domain person re-identification," 2019, *arXiv:1812.11369*.

[14] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May 2019, pp. 3146–3154.

[15] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.

[16] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person ReID," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11744–11752.

[17] G. Wang et al., "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6449–6458.

[18] X. Zhang, Y. Yan, J.-H. Xue, Y. Hua, and H. Wang, "Semantic-aware occlusion-robust network for occluded person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2764–2778, Jul. 2021.

[19] Z. Ma, Y. Zhao, and J. Li, "Pose-guided inter- and intra-part relational transformer for occluded person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1487–1496.

[20] K. Zheng, C. Lan, W. Zeng, J. Liu, Z. Zhang, and Z.-J. Zha, "Pose-guided feature learning with knowledge distillation for occluded person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia (ACM MM)*, 2021, pp. 4537–4545.

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

[22] 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. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8450–8459.

[23] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2898–2907.

[24] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4678–4686.

[25] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7073–7082.

[26] Y. Sun et al., "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 393–402.

[27] H. Luo, W. Jiang, X. Fan, and C. Zhang, "STNReID: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2905–2913, Nov. 2020.

[28] C. Szegedy et al., "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[29] Z. Wang, S. Zheng, M. Song, Q. Wang, A. Rahimpour, and H. Qi, "advPattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8341–8350.

[30] H. Wang, G. Wang, Y. Li, D. Zhang, and L. Lin, "Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 342–351.

[31] X. Wang, S. Li, M. Liu, Y. Wang, and A. K. Roy-Chowdhury, "Multi-expert adversarial attack detection in person re-identification using context inconsistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15097–15107.

[32] J. Li, Z. Du, L. Zhu, Z. Ding, K. Lu, and H. T. Shen, "Divergence-agnostic unsupervised domain adaptation by adversarial attacks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8196–8211, Sep. 2021, doi: 10.1109/TPAMI.2021.3109287.

[33] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 1–11.

[34] H. Tang, C. Yuan, Z. Li, and J. Tang, "Learning attention-guided pyramidal features for few-shot fine-grained recognition," *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108792.

[35] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[36] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.

[37] H. Tan, X. Liu, B. Yin, and X. Li, "MHSA-Net: Multi-head self-attention network for occluded person re-identification," 2020, *arXiv:2008.04015*.

[38] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, and W. Jiang, "SCPNet: Spatial-channel parallelism network for joint holistic and partial person re-identification," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Cham, Switzerland: Springer, 2018, pp. 19–34.

[39] D. Gray, S. Brennan, and T. Hai, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. 10th IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, no. 5. Rio de Janeiro, Brazil: IEEE, Oct. 2007 pp. 1–7.

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

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

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

[43] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 13001–13008.

[44] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.

[45] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

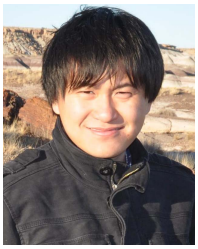
[46] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3219–3228.

[47] Y. Ge et al., "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. 32nd Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1–12.
[48] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 480–496.
[49] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5098–5107.
[50] H. Tan, X. Liu, Y. Bian, H. Wang, and B. Yin, "Incomplete descriptor mining with elastic loss for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 160–171, Jan. 2022, doi: 10.1109/TCSVT.2021.3061412.
[51] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Feature completion for occluded person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4894–4912, May 2022, doi: 10.1109/TPAMI.2021.3079910.
[52] Z. Ma, Y. Zhao, and J. Li, "Pose-guided inter- and intra-part relational transformer for occluded person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1487–1496.
[53] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.
[54] L. He and W. Liu, "Guided saliency feature learning for person re-identification in crowded scenes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 357–373.
[55] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.
[56] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2109–2118.
[57] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.
[58] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 365–381.
[59] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, and N. Zheng, "Discriminative feature learning with foreground attention for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4671–4684, Dec. 2019.
[60] 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, Oct. 2018.
[61] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.
[62] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 2013–2025, 2020.
[63] C. Wan, Y. Wu, X. Tian, J. Huang, and X.-S. Hua, "Concentrated local part discovery with fine-grained part representation for person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1605–1618, Jun. 2020.
[64] J. Miao, Y. Wu, and Y. Yang, "Identifying visible parts via pose estimation for occluded person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4624–4634, Sep. 2022, doi: 10.1109/TNNLS.2021.3059515.
[65] Y. Gong, L. Huang, and L. Chen, "Person re-identification method based on color attack and joint defence," 2021, *arXiv:2111.09571*.
[66] H. Jin, S. Lai, and X. Qian, "Occlusion-sensitive person re-identification via attribute-based shift attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2170–2185, Apr. 2022, doi: 10.1109/TCSVT.2021.3088446.
[67] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 402–419.

**Shujuan Wang** received the B.S. and Ph.D. degrees from Beijing Jiaotong University in 2007 and 2013, respectively. From 2010 to 2011, she was a Visiting Scholar at the Department of Electrical Engineering at Columbia University, NY, USA. She is an Associate Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology. Her research interests include deep learning and its application, Internet of Vehicles, vehicular communications, and intelligent transportations.

**Run Liu** received the B.E. degree from the College of Information Engineering and Automation, Kunming University of Science and Technology, China, in 2019, and the master's degree from the College of Information Engineering and Automation, Kunming University of Science and Technology, China, in 2022. His research interests include machine learning and computer vision.

**Huafeng Li** received the M.S. degree in applied mathematics and the Ph.D. degree in control theory and control engineering major from Chongqing University in 2009 and 2012, respectively. He is currently an Associate Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His research interests include image processing, computer vision, and information fusion.



**Guanqiu Qi** received the Ph.D. degree in computer science from Arizona State University in 2014. He is currently an Assistant Professor with the Computer Information Systems Department, State University of New York at Buffalo State. His research interests include deep learning, machine learning, and image processing. He also spans many aspects of software engineering, such as software-as-a-service (SaaS), testing-as-a-service (TaaS), big data testing, combinatorial testing, and service-oriented computing.

**Zhengtao Yu** received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language process, image processing, and machine learning.