Human Co-Parsing Guided Alignment for Occluded Person Re-identification

Shuguang Dou, Cairong Zhao, Xinyang Jiang, Shanshan Zhang, *Memeber, IEEE*, Wei-Shi Zheng, Wangmeng Zuo, *Senior Member, IEEE*

***Abstract*—Occluded person re-identification (ReID) is a challenging task due to more background noises and incomplete foreground information. Although existing human parsing-based ReID methods can tackle this problem with semantic alignment at the finest pixel level, their performance is heavily affected by the human parsing model. Most supervised methods propose to train an extra human parsing model aside from the ReID model with cross-domain human parts annotation, suffering from expensive annotation cost and domain gap; Unsupervised methods integrate a feature clustering-based human parsing process into the ReID model, but lacking supervision signals brings less satisfactory segmentation results. In this paper, we argue that the pre-existing information in the ReID training dataset can be directly used as supervision signals to train the human parsing model without any extra annotation. By integrating a *weakly supervised human co-parsing* network into the ReID network, we propose a novel framework that exploits shared information across different images of the same pedestrian, called the Human Co-parsing Guided Alignment (HCGA) framework. Specifically, the human co-parsing network is weakly supervised by three consistency criteria, namely global semantics, local space, and background. By feeding the semantic information and deep features from the person ReID network into the guided alignment module, features of the foreground and human parts can then be obtained for effective occluded person ReID. Experiment results on two occluded and two holistic datasets demonstrate the superiority of our method. Especially on Occluded-DukeMTMC, it achieves 70.2% Rank-1 accuracy and 57.5% mAP.**

***Index Terms*—person re-identification; image co-segmentation; human parsing.**

I. Introduction

PERSON re-identification [1] (ReID) aims to match a target pedestrian with non-overlapping cameras. In recent years, person ReID has been a research focus due to its potential application values in intelligent security and video surveillance [2]. However, most person ReID methods focus on holistic person ReID and ignore the occlusion problem. Consequently, directly matching two occluded pedestrian images without alignment can dramatically increase the difficulty of person ReID.

To solve the occlusion problem, various alignment-based methods have been proposed, which can be roughly divided into part-level [3], [4], [5], key point-level [6], [7], [8] and pixel-level [9], [10], [11], [12], [13]. Although existing pixel-level ReID methods can tackle this problem with semantic alignment at the finest level, those methods do not significantly outperform other alignment methods. This is because the performance of the pixel-level methods heavily relies on the human part segmentation model (or human parsing model). As shown in Figure 1(a) and (b), according to the supervised way of acquiring semantic information in ReID images, the existing pixel-level alignment methods can be divided into two types:

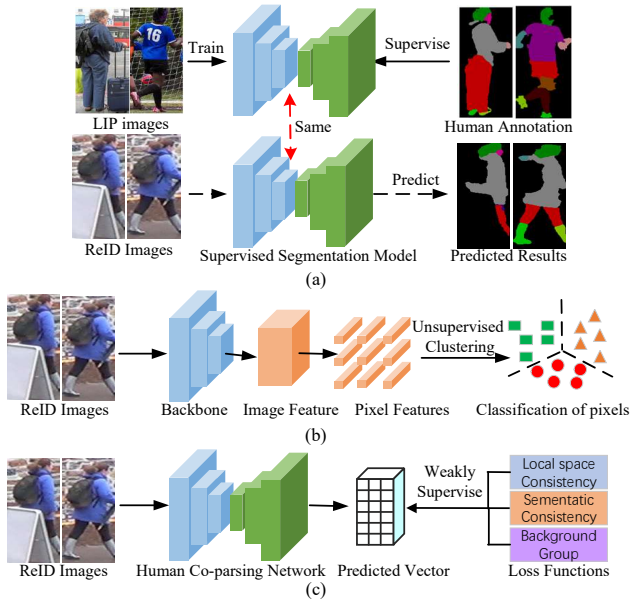

Fig. 1. Different supervised ways of acquiring semantic information in ReID images. (a) Supervised pixel-level alignment-based method: using a segmentation model supervised by pixel-wise labels of human annotation; (b) Unsupervised pixel-level alignment-based method: using the unsupervised clustering method with ReID feature map as input; (c) Weakly-supervised Human Co-parsing: Our method uses the pixel vicinity, semantic consistency, identity label as weakly supervised signals to train the segmentation model.

This work was supported by National Natural Science Fund of China (62076184, 61976158, 61976160, 62076182, 62276190), in part by Fundamental Research Funds for the Central Universities and State Key Laboratory of Integrated Services Networks (Xidian University), in part by Shanghai Innovation Action Project of Science and Technology (20511100700) and Shanghai Natural Science Foundation (22ZR1466700). (Corresponding author: Cairong Zhao.)

Shuguang Dou and Cairong Zhao are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: 2010504@tongji.edu.cn, zhaocairong@tongji.edu.cn).

Xinyang Jiang is with the Microsoft Research Asia (Shanghai), Shanghai 200232, China. (e-mail: xinyangjiang@microsoft.com)

Shanshan Zhang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: shanshan.zhang@njust.edu.cn).

Wei-Shi Zheng is with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, China (e-mail: wszheng@ieee.org).

Wangmeng Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: cswmzuo@gmail.com)

supervised methods and unsupervised methods.

The supervised pixel-level alignment-based methods use the pose estimation or human parsing model supervised by human annotation to obtain the pixel-level alignment information. For example, Huang *et al.* [10] use a human parsing model trained on COCO [14] to obtain pseudo pixel-wise labels of ReID images. Based on the pseudo-pixel-wise labels, a multi-task framework is proposed to train person ReID and human parsing in a shared backbone. However, because of the high cost of pixel-wise annotation, only segmentation models trained in other domains can be used, resulting in inaccurate segmentation results. Besides, when multiple pedestrians appear in an image, the supervised pixel-level alignment-based methods extract the pixel-level alignment information of irrelevant pedestrians as well, which may make the ReID task more difficult.

Different from the above supervised methods, the unsupervised pixel-level alignment-based methods use the unsupervised clustering method with the ReID feature map as input. For example, the Identity-guided Semantic Parsing [13] (ISP) designs the cascaded clustering on feature maps to generate pseudo-pixel-wise labels where cascade clustering consists of two k-means classifiers. However, using traditional clustering methods to obtain semantic information is sub-optimal due to lacking any supervision signals.

Instead of using extra cross-domain annotations, we train the human parsing model for ReID in a weakly supervised fashion. Inspired by co-part segmentation [15], we consider three characteristics of good human parsing and encode that prior knowledge into the loss function to train a weakly supervised segmentation network on ReID data. Specifically, as shown in Figure 1(c), we consider three desirable constraints of human parsing: (1) *Local space consistency*: Pixels in local space should be predicted as the same label. (2) *Semantic consistency*: Pixels with the same semantic in different images should be predicted as the same label. (3) *Background group*: The background of all images should be grouped into the same label.

We argue that the information that already exists in the ReID dataset can be utilized to achieve the above consistency constraint. For local space consistency, we assume that neighboring pixels should have the same semantic label. For semantic consistency, we take advantage of the information that the same ID has a similar appearance in different ReID images. For the background group, we utilize intra-camera view similarity in ReID images to separate the foreground from the background. The above information has not been fully exploited in the previous works. Since the proposed method cooperatively parses ReID images of the same ID together, we call our segmentation network Human Co-parsing Network.

The proposed method is a multi-task framework that jointly trains a human co-parsing network (HCNet) and a person ReID network (PRNet). 1) **HCNet**: a set of images with the same ID are firstly fed as a batch into an encoder and a decoder to generate prediction vectors. To satisfy the criteria, three consistency losses are designed to maximize the similarity of prediction vectors at three different levels. After training, common pedestrians and their personal belongings are separated from the background. However, the categories denoted by the predicted labels are unknown. To solve the problem, we design a center prior label reassignment (CPLR) to translate the unique predicted labels into the background and human parts. Finally, the co-parsing result is used as the pseudo-GT for the human parsing head of the person ReID network. When compared to recent pixel-level alignment-based methods [9], [10], [13], human co-parsing is relatively robust to occlusion. Both occluded objects and irrelevant pedestrians usually only occur in one image or one camera view and are not shared objects in all images. Therefore, human co-parsing treats the occluded objects and irrelevant pedestrians as background. 2) **PRNet** shares the backbone with HCNet and contains a human parsing head and an aligned person ReID head. The segmentation results predicted by HCNet are used as the pseudo-ground-truth to supervise the human parsing head. To refine the co-parsing results, we design Guided Alignment Module (GAM) by reducing the uncertainty of the segmentation prediction. On the one hand, GAM ignores pixel features with low confidence during training. On the other hand, GAM enhances pixel features with high confidence during testing.

The main contributions of this paper are summarized as follows:

1) We propose a novel Human Co-parsing Guided Alignment (HCGA) framework that alternately trains the human co-parsing network and the ReID network, where the human co-paring network is trained in a *weakly supervised manner* to obtain paring results without any extra annotation.
2) For the human co-paring network, we design three novel loss functions, namely local space consistency, semantic consistency, and background group, that satisfy the desirable constraint of human parsing.
3) For the ReID network, we propose a guided alignment module that reduces the uncertainty of the parsing prediction by ignoring pixel features with low confidence in the foreground during training and enhancing pixel features with high confidence during testing.
4) We conduct extensive experiments to demonstrate that the proposed method achieves superior performance on two occluded datasets—Occluded-DukeMTMC [16] and Occluded-REID [17], and competitive performance on two holistic datasets—Market-1501 [18] and CUHK03-NP [19], [20].

The remainder of the paper is organized as follows: In Section II, we review related work on person ReID and deep clustering methods. Section III details the proposed human co-parsing guided alignment framework. Sections IV and V present the comparison and evaluation of experimental results and visualization. Finally, conclusions are drawn in Section VI.

II. Related Work

The proposed HCGA focuses on person ReID, image co-segmentation, and deep clustering. Therefore, we briefly review the related work in this section.

A. Person Re-identification

Deep learning has achieved significant success in the task of person re-identification in recent years. Deep learning-based approaches integrate representation learning and metric learning into a unified network framework for optimization. In the existing deep learning-based methods, attention-based and part-level alignment-based methods are the two dominant approaches. The part-level alignment-based methods [4], [21], [3], [22] divide the pedestrian image into several regions or locate discriminative parts from the human body; The attention-based methods [23], [24], [25] guide the model to pay more attention to discriminative regions. Besides the above traditional directions, unsupervised domain adaptation, cross-modal and label noise learning have also attracted the attention of researchers. For example, Zhou *et al.* [26] propose a multi-feature fusion with adaptive graph learning for unsupervised Re-ID. Wang *et al.* [27] propose Attentive Waveblock (AWB) that can be integrated into the dual networks of mutual learning for depressing noise in the pseudo-labels. Ye *et al.* [28] propose a novel dynamic tri-level relation mining (DTRM) for visible infrared re-identification. To learn a robust ReID model, an online co-refining (CORE) [29] framework is proposed to distill the knowledge between different works. Most of those works have focused on the holistic ReID problem and do not consider the occluded problem.

Recently, some works explore the Vision Transformer for the ReID task. He *et al.* [30] first exploit pure Transformer for the ReID task and propose a Transformer-based Object Re-identification (TransReID) method. TransReID introduces the side information embedding to encode different side information and proposes the jigsaw patches module to utilize the stripe-based idea. Zhu *et al.* [31] propose an online Auto-Aligned Transformer (AAformer) to adaptive assign patch embedding of the same semantics to the same part token in the running time. AAformer learns part features and achieves part alignment by self-attention. Compared with the existing CNN-based methods, the transformer-based method is more robust to occlusion.

B. Occluded Person Re-identification

The occluded problem often occurs in real scenes. For example, the target pedestrian may be occluded by irrelevant pedestrians in crowded scenes. Zhuo *et al.* [17] firstly defines the occluded person ReID problem and proposes an Attention Framework of Person Body (AFPB) for the occluded problem. Recently, some pose estimation and human parsing-based methods are proposed to cope with the occluded ReID problem. Miao *et al.* [16] utilize a pose estimation model to extract useful information from the occluded images and guide the model to focus on non-occluded regions. Gao *et al.* [6] propose a pose-guide visible part matching method to fuse local features with visual scores. Wang *et al.* [7] firstly extract semantic local features by a pose estimation model and propose an adaptive direction graph convolution layers to learn relation and a cross-graph embedded-alignment layer to predict similarity score. Zhao *et al.* [32] propose a novel incremental generation of occlusion against suppression (IGOAS) network. IGOAS first generates easy-to-hard occlusion data by the incremental generation of occlusion blocks, and then suppresses the generated occlusion regions by adversarial suppression branches. Li *et al.* [33] first explores the transformer encoder-decoder structure for Occluded ReID and proposes a Part-Aware Transformer (PATrans) to learn part prototypes. PATrans designs part diversity and part discriminability to achieve robust human part discovery. Jia *et al.* [34] propose a disentangled representation learning network (DRL-Net) that handles occluded re-ID without requiring strict person image alignment.

However, most of those occluded ReID methods require extra annotation for fine-grained alignment. Besides, the performance of the occluded person ReID methods on holistic datasets has been relatively lower performance. The holistic and occluded person ReID tasks are not contradictory to each other. Different from the above occluded ReID methods, our proposed method is valid for both tasks. Our previous work IGOAS used data augmentation and attention mechanisms to address the occlusion problem. Different from IGOAS, this work is motivated to address the lack of manually annotated semantic information in existing ReID datasets. In this paper, we propose a weakly supervised approach to obtain semantic information in ReID images to guide ReID network at pixel-level alignment.

C. Deep clustering

Deep clustering has progressed in recent years due to the powerful feature extraction capabilities of deep convolutional neural networks. Ji *et al.* [35] propose an invariant information clustering to maximize the mutual information of sample pairs. Kim *et al.* [36] propose an unsupervised image segmentation method based on differentiable feature clustering. The method exploits spatial continuity to replace the constraints of fixed segment boundaries, such as over-segmentation. Chen *et al.* [37] propose a novel method that exploits semi-supervised learning for the imbalanced problem. Different from the above deep clustering methods, Li *et al.* [38] use the label as a special representation and propose contrastive clustering (CC) based on contrastive learning. CC maximizes the similarity of sample pairs at the instance level and the cluster level. Inspired by the above works, we train an image segmentation network by jointly maximizing the similarity of predicted vectors at three different levels.

D. Image Co-segmentation

To the best of our knowledge, the concept related to co-parsing was first introduced in clothing co-parsing [39]. Some other problems that are similar to Human Co-parsing are Instance Co-segmentation [40] and Co-part Segmentation[15]. The problem to be solved by Co-part Segmentation is part segmentation between different object instances for collections of images with only one object class. Instance co-segmentation aims to identify all instances of objects in a set of images that together contain a specific class and segment each instance. Different from the above image co-segmentation task, human co-parsing is the ID-level part co-segmentation task rather than the class-level.

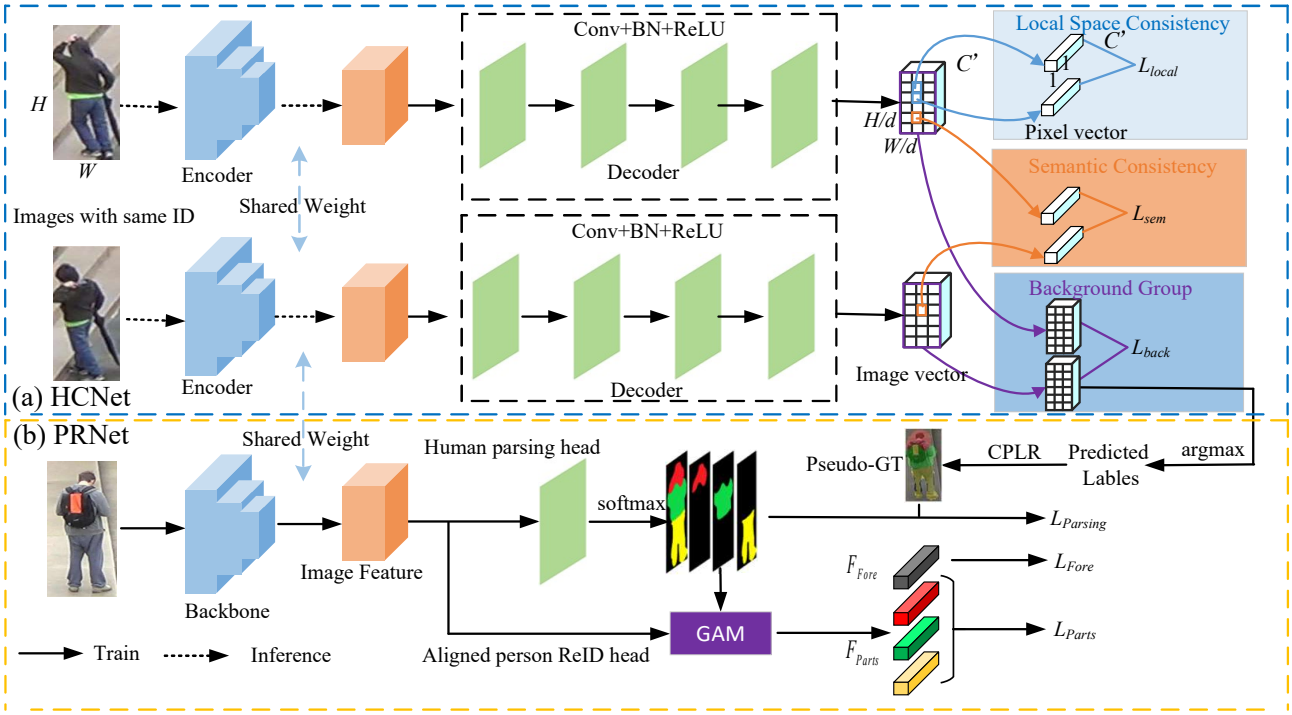

Fig. 2. The flowchart of the proposed HCGA framework. (a) Human Co-parsing network (HCNet). (b) Person ReID network (PRNet). The proposed HCGA consists of HCNet and PRNet. The two sub-networks share the backbone and are trained alternately until optimized to the best. Specifically, PRNet refines the parameters to make the difference between features with different semantics greater. HCNet yields better segmentation results based on the refined features to guide PRNet alignment at the pixel level.

III. Human Co-parsing Guided Alignment Framework

In this section, we first describe the overall framework. Second, we describe in detail the Human Co-Parsing network architecture and three proposed consistency losses. Third, we describe the person ReID network and the structure of the Guided alignment module. Finally, inference processing is described.

A. Overall Framework

As shown in Fig.2, the proposed HCGA framework consists of two sub-networks and the encoder of HCNet is the backbone of the PRNet. Each epoch of the training phase of HCGA consists of two steps: (1) a set of images of the same ID in the training set is fed as a batch into the HCNet. In the training phase of HCNet, the parameters of the Encoder are not updated. For each ID of the training set, we train a Decoder separately and output co-parsing results for all images of that ID at the end of training. (2) The co-parsing results are used as the pseudo-GT of the human parsing head of the PRNet. In the training, all parameters of PRNet are updated by backpropagation.

In the early stage of training, the segmentation effect of HCNet is not good. As PRNet keeps optimizing the parameters in Backbone, the difference between pixel features of foreground and background increases. Therefore, HCNet generates better segmentation results with iterative training. Guided by the semantic information obtained from HCNet, PRNet is robust to occlusion.

B. Human Co-parsing Network

Before elaborating on HCNet, the problem to be solved by the human co-parsing network is described as follows. Let $X = \{I_k(i,j) \in \mathbb{R}^3\}_{k=1}^N$ be a set of images with same ID where N denotes the number of images and i and j denote the position of the pixel in kth image I. Let $E : \mathbb{R}^3 \rightarrow \mathbb{R}^q$ be an encoder for feature extraction and $\{V_k(i,j) \in \mathbb{R}^q\}_{k=1}^N$ be the q-dimensional feature vectors of the pixel in row i and column j of the kth image. The output that the segmentation network requires is pixel-wise labels $\{Y_k(i,j) \in \mathbb{N}\}_{k=1}^N$. In the labels, 0 indicates the background and 1 to $C-1$ indicates human body parts and personal belongings and C indicates the total number of categories in multiple images. Let $D : \mathbb{R}^q \rightarrow \mathbb{N}$ denotes a decoder for mapping the feature vectors to labels. Collectively, the problem can be formulated as follows:

$$Y_k(i,j) = D(E(\{I_k(i,j)\}_{k=1}^N)) \tag{1}$$

The human co-parsing network aims to solve the weakly supervised image segmentation problem. Only $\{I_k(i,j)\}_{k=1}^N$ is known in Eq. (1) while D, E, and labels $\{Y_k(i,j)\}_{k=1}^N$ are unknown, thus solving this equation is challenging. To achieve co-parsing, we design a deep convolutional neural network with an encoder-decoder structure and three consistency losses to weakly supervise the segmentation network.

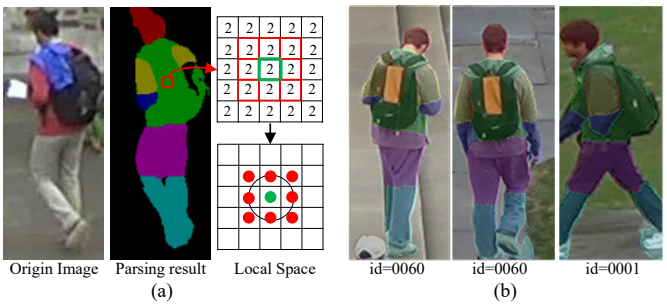

Fig. 3. Consistency information within and between images. (a) local space consistency. (b) Semantic consistency between images.

Network Architecture. HCNet is shown in Figure 2(a). We use a backbone as an encoder and design a decoder consisting of several convolutional layers. Three loss functions are designed to satisfy the criteria of parsing. Finally, the central prior-based label reassignment (CPLR) transforms the predicted labels into the background and human parts. The pixel values of the image set X are first normalized to [0,1] and then transported to the encoder. The parameters in the encoder are fixed when extracting the feature vectors $\{V_k(i,j)\}_{k=1}^{N}$, and the encoder is trained only in the Person ReID network. The feature vectors $\{V_k(i,j)\}_{k=1}^{N}$ are then fed into a decoder. The decoder consists of four basic blocks that are composed of a convolutional layer, a batch normalization layer [41], and a ReLU activation function. The last basic block contains a convolutional layer with $C'(C' >> C)$ convolutional kernels of size 1×1 and a BN function. Finally, the classification result of each pixel is output after the argmax function. Notably, the network output predicted labels are $\{Y'_k(i,j) \in \{0,1,\cdots C'-1\}\}_{k=1}^{N}$.

Local space consistency loss. As shown in Figure 3(a), when focusing only on one pedestrian image and its corresponding human parsing result, it is observed that a pixel has the same label as its neighboring pixels in the local space, i.e., local spatial consistency. However, how to guarantee the local space consistency of segmentation results. The LBP operator [42] is defined as the comparison of the grayscale values of 8 adjacent pixels within a 3×3 window, with the center pixel of the window as the threshold. Inspired by the LBP operator [42], we maximize the similarity of the prediction vectors of the center pixel and the neighboring pixels within a $R\times R$ window. In short, we propose the local space consistency loss which can be formulated as:

$$\mathcal{L}_{local} = \sum_{s=1}^{S} \|y_c - y_s\|_p \tag{2}$$

where $y_c \in \mathbb{R}^{1\times1\times C'}$ is the prediction vector of the center pixel, $y_s(s=1,\cdots,S)$ is the prediction vector of neighboring pixels from a $R\times R$ neighborhood, and $\|\cdot\|_p$ is p-norm. In this paper, R is set to 3.

Semantic consistency loss. A human can easily annotate pixels with the same semantics of different images as the same label, but it is difficult for the network to learn this semantic consistency without pixel-wise labels. As shown in Figure 3(b), the arms or legs of different ids have the same semantics but different colors and textures, but the human parts of the same ID have the same color and texture under different camera views. Therefore, for all images of the same id, the predicted vectors of pixels with the same semantics in the pedestrian region are similar and the network assigns the same label to similar predicted vectors. We assume that pixels with the same semantic in the pedestrian region are assigned the same label by the network. Based on the assumption, we propose the semantic consistency loss to maximize the similarity of pixels vectors with the same semantic across images, which can be formulated as:

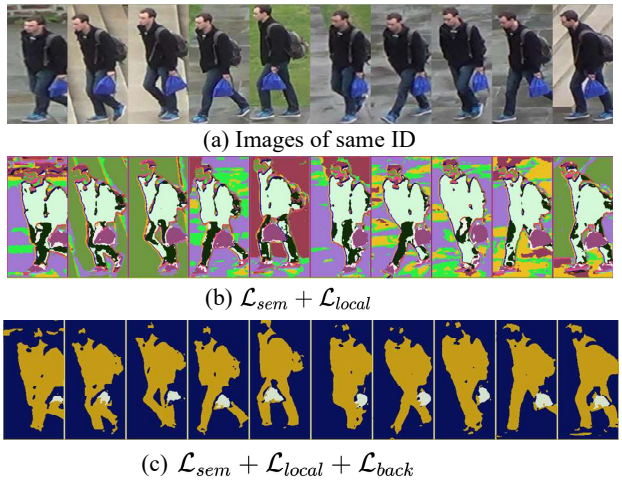

Fig. 4. Visualization of the co-paring results of HCNet trained directly on RGB images with the same ID. (a) Images of the same ID. (b) Co-paring results when only local consistency loss and semantic loss are used. (c) Co-paring results when three losses are used.

$$\mathcal{L}_{sem} = -\sum_{i=1}^{M} \log\left(\frac{\exp(y_{pl})}{\sum_{j=1}^{C'}\exp(y_j)}\right) \tag{3}$$

where M is the number of all pixels in the set of images with the same id, y is the prediction vector of pixels and pl is the pseudo-label obtained by the predicted vector passing through the argmax function.

Background group loss. Since the background is different across images, only local consistency loss and semantic consistency loss cannot guarantee that the background pixels are grouped into one class, as shown in Fig.4 (b). To achieve it, we mine the shared information in the ReID data. The lighting and pose of the ReID images with the same ID change considerably under different camera views. However, the ReID images are similar in the same camera view except for the background bias, i.e., intra-camera view similarity. Based on the intra-camera view similarity, we maximize the similarity of the prediction vectors of the neighboring images from the same camera view to eliminate background bias. In short, we propose a background group (BG) loss that can group the background pixels into one class. The formulation of BG loss is shown as follows:

$$\mathcal{L}_{back} = \sum_{k_1\neq k_2}^{N-1} \left\|y_{k_1} - y_{k_2}\right\|_p \tag{4}$$

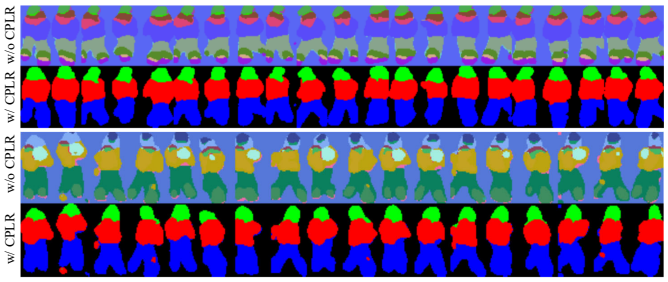

Fig. 5. Comparison of segmentation results without CPLR and with CPLR

where $y_k \in \mathbb{R}^{H/d \times W/d \times C'} (k = 1, \cdots, N)$ is the prediction vector of kth image with size $H \times W$, k_1th and k_2th images are the adjacent images under the same camera view. In this paper, the possible values of p are 1 and 2. The background group loss corrupts the other two consistency losses to some extent. To reduce this corruption, we designed the training abort mechanism. In the background group loss, the background bias provides a larger gradient compared to the pedestrian bias. After the background is essentially predicted as the same label by the network, we abort the training (details in section V-B).

CPLR. The weakly-supervised image segmentation network cannot output a specific category represented by each prediction label. Therefore, label reassignment is required on the co-parsing results. The CPLR divides the predicted labels into foreground and background based on prior knowledge that the foreground is generally in the center of the image while the background is mostly at the edge. According to the average height of the pixels of the unique predicted labels, the CPLR divides the foreground labels into $C-1$ human parts. The segmented image before and after CPLR processing are shown in Figure 5.

Optimization. At the initiation of network training, the parameters in the encoder E are initialized with the classification model pre-trained in ImageNet, while the parameters in the mapping function D are initialized with Kaiming uniform [43]. Since E and D are determined, the predicted label $Y'_k(i,j)$ can be obtained by forward propagation. The objective function of HCNet can be formulated as:

$$\mathcal{L}_{HCNet} = \mathcal{L}_{sme} + \lambda_L \mathcal{L}_{local} + \lambda_B \mathcal{L}_{back} \tag{5}$$

where λ_L and λ_B is the balance weights. During the iterative training process, the number of unique predicted labels continuously becomes less due to the constraint of Equation 5. Once the number of unique predicted labels is smaller than the minimum number min_L, the training process ceases and the co-parsing results are fed into the CPLR to generate the final segmentation result. The pseudocode for the weakly supervised human co-parsing methods is shown in Algorithm 1. Notably, θ_E is only updated by PRNet.

Algorithm 1 Weakly Supervised Human Co-parsing

Input: N images with same ID $X = \{I_k(i,j) \in \mathbb{R}^3\}_{k=1}^N$, minimum number of unique predicted labels min_L, learning rate μ
Output: Labels $\{Y_k(i,j) \in \mathbb{N}\}_{k=1}^N$

```
Initialize θ_E with pre-trained parameters in ImageNet
Initialize θ_D with Kaiming uniform
{V_k(i,j)}_{k=1}^N ← E({I_k(i,j)}_{k=1}^N)
repeat
    {Y'_k(i,j)}_{k=1}^N ← D({V_k(i,j)}_{k=1}^N)
    L_HCNet ← L_sme + λ_L L_local + λ_B L_back
    Update θ_D ← θ_D − μ ∂/∂θ_D L_HCNet
until Unique({Y'_k(i,j)}_{k=1}^N) < min_L
Get the predicted labels on the image boundaries boundary
labels, hists ← Unique(boundary, return_counts = True)
for label, hist in enumerate(labels, hists) do
    if hist > image_perimeter/2 then
        Reassign label to background label b_labels
    end if
end for
f_labels ← set(all_labels) - set(b_labels)
Reassign the foreground labels f_labels to C − 1 human parts based on the average height
return {Y_k(i,j)}_{k=1}^N =0
```

C. Person ReID network

The PRNet is based on ISP [13]. Using the backbone with shared weights to extract deep features, the human parsing head first generates C confidence maps by a convolution layer with C convolution kernels of size 1×1 and a softmax function. The human parsing loss is calculated by the prediction result of human parsing head and the pseudo-GT generated by HCNet. In the aligned person ReID head, the confidence map and depth features are fed into the GAM, whose structure is shown in Figure 5. The confidence map of the foreground CM_{fore} is obtained by summing $1st$ to Cth confidence map, which can be expressed by the following equation:

$$CM_{Fore} = \sum_{i=1}^{C} CM_i \tag{6}$$

In GAM, we treat the confidence maps of human parts CM_{part} differently during training and testing depending on the confidence (or probability values). Specifically, we suppress the pixel features with low confidence during training, which can be formulated as follows:

$$CM_{\text{part}}(e) = \begin{cases} e, e > \frac{1}{t} \\ 0, e \le \frac{1}{t} \end{cases} \tag{7}$$

In inference, we set the high confidence as a threshold to binary the confidence map.

$$CM_{\text{part}}(e) = \begin{cases} 1, e > 1 - \frac{1}{t} \\ 0, e \le 1 - \frac{1}{t} \end{cases} \tag{8}$$

where e is the element of confidence map and $t \in \mathbb{N}_+$ determine the threshold. The image features are multiplied with a confidence map of foreground CM_{Fore} and human parts CM_{part} at the element level to obtain the features of foreground F_{Fore} and human parts F_{part}.

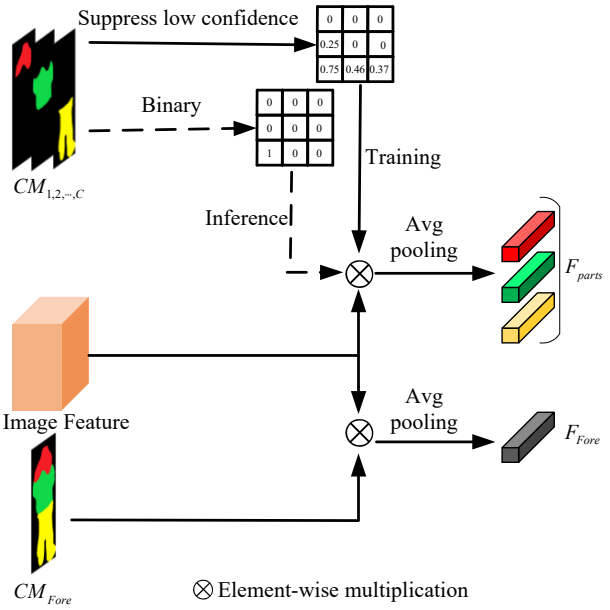

Fig. 6. The structure of Guided Alignment Module (GAM)

After the average pooling operation, the features of the image, foreground, and human parts are used to predict the ID with three different classifiers. The objective function of PRNet is formulated as:

$$\mathcal{L}_{PRNet} = \mathcal{L}_{Image} + \mathcal{L}_{Fore} + \mathcal{L}_{Parts} + \lambda_{hp}\mathcal{L}_{Parsing} \quad (9)$$

where $\mathcal{L}_{Image}$ is the triplet loss [44] and cross-entropy loss with label smoothing [45], $\mathcal{L}_{Fore}$ and $\mathcal{L}_{Parts}$ are cross-entropy loss with label smoothing, $\mathcal{L}_{\text{Parsing}}$ is pixel level cross-entropy loss, λ_{hp} is the balance weight.

D. Inference

In the inference phase, only PRNet is required. Therefore, HCNet does not add additional consumption during inference. For a pair of images (x_1, x_2), we use PRNet to get the whole image features, foreground, and human parts features of the image pair. Considering the large noise interference in the occluded image, we use only the foreground and human parts features to calculate the similarity of the image pair. For the distance of human parts features, similar to the recent occluded ReID methods [16], [10], [13], we only calculate the distance of shared visible parts.

$$sim(x_1, x_2) = \frac{\cos(F^{x_1}_{Fore}, F^{x_2}_{Fore}) + \frac{1}{2}\sum_{i=1}^{C-1} P_i \cos(F^{x_1}_{part_i}, F^{x_2}_{part_i})}{1 + SV} \quad (10)$$

where $\cos(\cdot)$ is cosine distance, P_i equal 1 if $part_i$ is shared visible part of the image pair else 0, and $SV \leq C$ is the number of share visible parts.

IV. Experiments

A. Implementation Details

Datasets. To verify that the proposed HCGA is effective for occluded and holistic ReID problems, we conduct experiments on two holistic and two occluded datasets. The details of the four data sets are as follows: 1) Occluded-DukeMTMC [16] consists of 15,618 training images, 2,210 occluded query images, and 17,661 gallery images. 2) Occluded-REID [17] contains 1000 holistic images and 1000 occluded images of 200 IDs. Half of Occluded-REID is used for training and the remaining half for testing. 3) CUHK03-NP [19], [20] uses a new testing protocol similar to Market-1501, which divides the dataset into a training set containing 767 pedestrians and a test set containing 700 pedestrians. 4) Market-1501 [18] contains 12,936 training images, 3368 query images, and 15,913 gallery images. Example images of the above four datasets are shown in Fig.7

Fig. 7. Example images from two occluded ReID datasets and two holistic ReID datasets

Training Details. The HCGA framework is implemented by Pytorch. All images of the training set are resized to 256×128 and augmented with random erasing[46], horizontal flipping, random cropping, and padding 10 pixels. For the HCNet, each batch is all images that contain the same ID. We set C' to 32 and min_L to 18. The balance weights λ_L and λ_B are 2 and 1. The optimizer of HCNet is SGD with a momentum of 0.9. The segmentation network is only trained for 32 epochs with a learning rate of 0.1. For the PRNet, all parameters are trained for 120 epochs with the Adam optimizer. The learning rate is 3.5e-4 and decays to 0.1 at 40 and 70 epochs. The batch size is 64 and the balance weight λ_{hp} is 0.1.

Evaluation Metrics. Following most works in person ReID, the Cumulative Matching Characteristic curves (CMC) at Rank-1 and Rank-5 and the mean average precision (mAP) are used in this paper to evaluate the performance of different person ReID methods. All experiments are implemented on NVIDIA RTX 3090 GPU and in the single query setting.

B. Experimental Results

Results on Occluded Datasets. As shown in Table I and Table II, we compare our method with 6 holistic person ReID methods: SVDNet [50], HACNN [24], DSR [47], PCB [3], SFR [48], MLFN [51], 7 state-of-the-art (SOTA) person occluded ReID methods: AFPB [17], Teacher-S [52], PGFA [16], REDA [46], HONet [7], ISP [13], MOS [49] and 3 transformer-based ReID method: PATrans [33], AAFormer [31], TransReID [30] and DRL-Net [34]. For the Occluded-Duke dataset, the key point-level alignment-based methods

TABLE I
COMPARISON WITH STATE-OF-THE-ART CNN-BASED METHODS ON OCCLUDED-DUKEMTMC (%)

Methods	References	Rank-1	Rank-5	Rank-10	mAP
HACNN [24]	CVPR18	34.4	51.9	59.4	26.0
DSR [47]	CVPR18	40.8	56.2	65.2	30.4
PCB [3]	ECCV18	42.6	57.1	62.9	33.7
SFR [48]	ArXiv18	42.3	60.3	67.3	32.0
PGFA [16]	CVPR19	51.4	68.6	74.9	37.3
HONet [7]	CVPR20	55.1	-	-	43.8
ISP [13]	ECCV20	62.8	78.1	82.9	52.3
IGOAS [32]	TIP21	60.1	-	-	49.4
MOS [49]	AAAI21	66.6	-	-	55.1
HCGA$_{w/o\ GAM}$ (Ours)		65.9	80.2	84.3	56.0
HCGA(Ours)		**70.2**	**83.3**	**87.0**	**57.5**

TABLE II
COMPARISON WITH STATE-OF-THE-ART TRANSFORMER-BASED METHODS ON OCCLUDED-DUKEMTMC (%)

Methods	References	Rank-1	Rank-5	Rank-10	mAP
PATrans [33]	CVPR21	64.5	-	-	53.6
AAFormer [31]	ArXiv21	67.0	81.5	86.1	58.2
TransReID [30]	ICCV21	66.4	-	-	**59.2**
DRL-Net [34]	TMM22	65.0	79.3	83.6	50.8
HCGA(Ours)		**70.2**	**83.3**	**87.0**	57.5

TABLE III
COMPARISON WITH STATE-OF-THE-ART CNN-BASED METHODS ON OCCLUDED-REID DATASETS (%)

Methods	References	Rank-1	Rank-5
SVDNet [50]	ICCV17	63.1	85.1
MLFN [51]	CVPR18	64.7	87.7
PCB [3]	ECCV18	66.6	89.2
REDA [46]	AAAI20	65.8	87.9
AFPB [10]	ICME18	68.1	88.3
Teacher-S [52]	Arxiv18	73.7	92.9
ISP [13]	ECCV20	86.2	95.4
IGOAS [32]	TIP21	81.1	91.6
HCGA$_{w/o\ GAM}$ (Ours)		87.2	95.6
HCGA(Ours)		**88.0**	**96.0**

TABLE IV
COMPARISON WITH STATE-OF-ART CNN-BASED METHODS ON MARKET-1501. THE SECOND ROW IS THE PART-LEVEL ALIGNMENT-BASED METHODS. THE THIRD ROW IS THE KEY POINT-LEVEL ALIGNMENT-BASED METHODS. THE FOURTH ROW IS PIXEL-LEVEL ALIGNMENT-BASED. THE FIFTH ROW IS THE ATTENTION-BASED METHODS. (%)

Methods	References	Market-1501		
		Rank-1	Rank-5	mAP
PCB+RPP [3]	ECCV18	92.3	97.5	77.4
MGN [53]	MM18	95.7	-	86.9
Relation Net [54]	AAAI20	95.2	-	88.9
PABR [55]	ECCV18	91.7	96.9	85.0
PGFA [16]	ICCV19	91.2	-	76.8
HONet [7]	CVPR20	94.2	-	84.9
SPReID [11]	CVPR18	92.5	-	81.3
P^2-Net [56]	ICCV19	95.2	98.2	85.6
FPR [57]	CVPR19	95.4	-	86.6
ISP [13]	ECCV20	95.3	98.6	88.6
MPN [58]	PAMI 22	96.3	-	89.4
MPN* [58]	TPAMI 22	**96.4**	-	**90.1**
DuATM [59]	CVPR18	91.4	97.1	76.6
MHN-6 [23]	ICCV19	95.1	98.1	85.0
SCSN [60]	CVPR20	95.7	-	88.5
RGA-SC [61]	CVPR20	96.1	-	88.4
HCGA$_{w/o\ GAM}$ (Ours)		95.3	98.4	87.8
HCGA(Ours)		95.2	98.2	88.4

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON CUHK03-NP (%)

Methods	Detected		Labeled	
	Rank-1	mAP	Rank-1	mAP
PCB+RPP[3]	63.7	57.5	-	-
MGN [53]	66.8	66	68	67.4
Relation Net [54]	74.4	69.6	77.9	75.6
MHN-6 [23]	71.7	65.4	77.2	72.4
Auto-ReID [62]	73.3	69.3	77.9	73
ISP[13]	75.2	71.4	76.5	74.1
P^2-Net [56]	74.9	68.9	78.3	73.6
BDB+Cut[63]	76.4	73.5	79.4	76.7
DSA-reID[64]	78.2	73.1	78.9	75.2
Pyramid[65]	78.9	74.8	78.9	76.9
MPN[58]	**83.4**	**79.1**	**85.0**	**81.1**
HCGA(Ours)	76.9	73.2	78.3	75.8

are about 10% higher than the holistic ReID methods in Rank-1. The pixel-level alignment-based method ISP [13] has significantly higher performance than them. Compared with the second-best CNN-based method MOS, HCGA improved by 3.6% in Rank-1 and 2.4% in mAP. Compared with the Transformer-based approach, HCGA shows competitive performance on both Rank-1, Rank-5, and Rank-10. For the Occluded-ReID dataset, two ReID occluded methods REDA [46] and AFPB [17] and two holistic methods MLFN [51] and PCB [3] achieve similar performance. The reason may be that the Occluded-ReID dataset is relatively small. Similarly, our method improves 1.8% in Rank-1 compared with the second-best method ISP. Notably, for a fair comparison, we do not list the performance of HONet and PGFA on Occluded-ReID. This is because HONet and PGFA use a different dataset division method from AFPB, which proposes the Occluded-REID dataset.

Results on Holistic Datasets. As shown in Table IV, we compare the proposed method with three part-level alignment-based methods: PCB+RPP [3], MGN [53], Relation Net[54], three Key point-level alignment-based methods: PABP [55], PGFA [16], HONet [7], four pixel-level alignment-based methods: SPReID [11], P^2-Net [56], FPR [57], ISP [13], four attention-based methods: DuATM [59], MHN-6 [23], SCSN [60], RGA-SC [61]. The key point-level alignment-based methods have slightly lower performance than the other methods on two holistic datasets. These methods that directly use pre-trained human pose estimation models may generate similar problems as mentioned in the section I. The pixel-level alignment-based method MPN [54] achieves the best performance of 96.4% in Rank-1 on the Market-1501 dataset and 85.0% Rank-1 on CUHK03-NP labeled dataset. However, MPN uses two additional types of information: human paring [66] and human segmentation [12]. Compared with the state-of-the-art methods in different directions, our method still achieves comparable performance.

C. Ablation Study

Analysis of the choice of C. The hyperparameter C is predefined which affects the granularity of the parsing of the

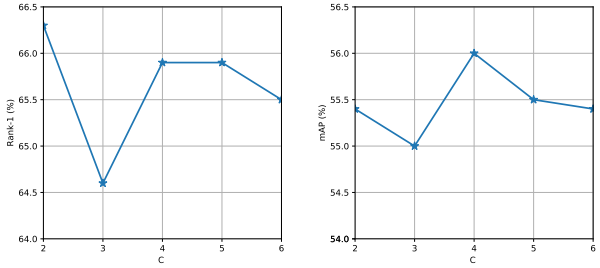

Fig. 8. Analysis of C for HCGA without GAM on Occluded-Duke dataset.

TABLE VI
ANALYSIS OF THE EFFECT OF SEGMENTATION ON REID PERFORMANCE (%)

Methods	CUHK03-Labeled		CUHK03-Detected		Occluded-Duke	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	71.9	68.5	67.6	64.7	57.8	49.7
+Extra info	73.3	71.9	72.2	69.6	61.2	50.2
+ISP	76.5	74.1	**75.2**	71.4	62.8	52.3
+HCNet	**77.4**	**75.7**	74.7	**71.9**	**65.9**	**56.0**

PRNet. Therefore, we analyze the impact of the choice of C for HCGA without GAM on the ReID task in this part. As shown in Figure 8, HCGA is robust to C. The best mAP and good Rank-1 of Occluded-Duke are obtained when the value of C is 4. It is reasonable to divide pedestrians into three parts because the main features of pedestrians are the head, upper body, and lower body. Therefore, we choose $C = 4$ for all datasets.

Analysis of the effect of segmentation on ReID performance. We analyze the effect of segmentation on ReID performance in this part. For the comparison with similar methods, we set the HRNet-W32 [67] as the baseline model following the ISP [13]. "+Extra info" denotes using human parsing result generated by a pre-trained human parsing model SCHP [68] as the pseudo-GT of the human parsing part. "+ISP" denotes using the clustering result generated by cascaded clustering [13] as the pseudo-GT. "+HCNet" denotes using the co-parsing result generated by HCNet as the pseudo-GT. As shown in Table VI, the cascaded clustering that uses the traditional clustering algorithm to classify the pixels of all images of the same ID is superior to the direct use of extra semantic information. Human Co-parsing, learning three constraints, has the highest performance in both datasets.

Analysis of the loss function. In this part, we analyze the three consistency losses. As shown in Table VII, when using only $\mathcal{L}_{sem}$, our method drops 4.3% in Rank-1 and 3.1% in mAP on Occluded-Duke datasets. When using $\mathcal{L}_{sem}$ and $\mathcal{L}_{local}$, our method drops 2.6% in Rank-1 and 3.0% in mAP on Occluded-Duke datasets. Similarly, there is more performance degradation on the Occluded-Duke dataset when $\mathcal{L}_{sem}$ and $\mathcal{L}_{back}$ are used. And on the holistic dataset DukeMTMC-reID, there is less performance degradation. One possible reason is that for the holistic dataset, the distance of features of the entire image is calculated besides the foreground features and shared-visible human parts features. The experimental result shows that optimizing three consistency losses simultaneously leads to better performance.

TABLE VII
ANALYSIS OF THE LOSS FUNCTION (%)

$\mathcal{L}_{sem}$	$\mathcal{L}_{local}$	$\mathcal{L}_{back}$	Occluded-Duke	
			Rank-1	mAP
✓	×	×	61.6	52.9
✓	✓	×	63.3	53.0
✓	×	✓	62.2	52.3
✓	✓	✓	**65.9**	**56.0**

TABLE VIII
ANALYSIS OF DIFFERENT FEATURE AND THE THRESHOLD ON OCCLUDED-DUKE. THE SECOND ROW DOES NOT USE GAM (%)

	Rank-1	Rank-5	Rank-10	mAP
F	65.7	79.8	84.3	55.2
P	65.3	79.7	83.8	55.6
F+P	**65.9**	**80.2**	**84.3**	**56.0**
I+F	63.2	77.6	82.1	53.7
I+F+P	64.6	79.0	83.1	54.9
GAM(t=2)	66.2	79.8	84.2	54.7
GAM(t=3)	68.0	82.5	**87.2**	55.5
GAM(t=4)	**70.2**	**83.3**	87.0	**57.5**

TABLE IX
ANALYSIS OF THE EFFECT OF NEIGHBORHOOD R ON OCCULDED-DUKE

Neighborhood	Rank-1	Rank-5	Rank-10	mAP
$R \times R = 5 \times 5$	68.8	81.8	85.3	51.9
$R \times R = 3 \times 3$	70.2	83.3	87.0	57.5

Analysis of different feature and the threshold under occlusion. We analyze the effect of using different features under occlusion. I, F, and P represent image features, foreground features, and human parts features respectively. The experiments in the second row of Table VIII do not use GAM, i.e., we directly multiply the confidence of human parts CM_{part} with the image features to obtain the human parts features F_{part} without suppressing the low confidence of pixel features and enhancing the high confidence. First, using F+P has a better performance compared to using only F or P. Second, using image features for occluded person ReID leads to performance degradation compared to using only foreground features or human parts features, because the image features contain a large amount of background noise in the occluded scenes. Finally, for GAM, the threshold of the confidence map is 0.5 for both training and inference when t=2, and the performance improves slightly on Rank-1 compared to F+P. When t>2, the performance on Rank-1 rises significantly. The GAM performs best for high-confidence feature enhancement while suppressing low confidence.

Analysis of the effect of neighborhood R. In this part, we analyze the impact of neighborhood R. As shown in Table IX, when we expand the range of R, the mAP decreases substantially. Since the local spatial consistency loss operates on the reduced-dimensional features rather than directly on the pixels, setting $R = 5$ makes the effect of clustering coarse. Therefore, we simply set $R = 3$ to get good results.

Analysis of the Encoder. We analyze the impact of different methods for the initialization of the encoder. As shown in Table X, we compare HRNet [67] with ResNet [69]. For ResNet, we use upsampling layer to linearly interpolate the

TABLE X
ANALYSIS OF THE ENCODER. "PARAM" DENOTES THE PARAMETERS OF THE ENCODER AND "FM SIZE" DENOTE THE OUTPUT SIZE OF THE FEATURE MAP BY THE ENCODER.

Encoder	Param	FM Size	Occluded-Duke		CUHK03-Labeled	
			Rank-1	mAP	Rank-1	mAP
HRNet-W32	28.5M	64×32	70.2	57.5	78.3	75.8
ResNet50	28.1M	16×8	61.0	45.9	70.6	67.4

TABLE XI
AVERAGE TRAINING TIME OF EACH EPOCH FOR TWO SUB-NETWORKS. "S" DENOTES SECOND.

Networks		CUHK03-NP	Market	Occluded-Duke
HCNet	Computer feature	46s	69s	78s
	Co-paring	91s	102s	119s
PRNet		258s	323s	390s

16×8 feature map to 64×32 which is the same as HRNet. The performance of HRNet is better than that of ResNet, although the parameters of both are similar. As mentioned by ISP [13], segmentation or co-paring has a high requirement for the resolution of the feature maps and needs more semantic information.

D. Runtimes

We report the running time of each part of the entire framework in Table XI. For HCNet, we have to compute the features by inference first, and then send the features to Decoder for training. Compared with the time to train PRNet, inferring features is faster. Since all images of each ID are trained independently, we speed up the co-paring process by multiple processes. Compared to training only PRNet, it takes about 1.5 times longer to train the whole framework. However, the HCNet does not need to be trained to the maximum epoch, the co-paring effect of the HCNet is good enough when training to certain epochs, so we can shorten the training time by stopping the HCNet early.

V. VISUALIZATION

A. Comparison with existing pixel-level alignment-based methods

For occluded person ReID problems, pedestrian occlusion is common but difficult to solve, especially for pixel-level alignment-based methods. Human Co-parsing aims to segment the common objects from images, while pedestrian occlusion often occurs in only one camera view, i.e., irrelevant pedestrians are not the common objects in multi-camera views. As shown in Figure 9, HCGA segment only the common pedestrian from the images compared with SCHP and ISP. In the case where three pedestrians are occluded from each other, HCGA also minimizes the interference of irrelevant pedestrians. Although HCGA enables the network to focus only on the target pedestrians in the face of occlusion, the results of Co-paring are rougher at the edge part of the person. It may be caused by the loss of local spatial consistency, which forces neighboring pixels to be grouped into one category. Based on the above phenomenon, we propose the guided alignment module to ignore the low-confidence pixel features to refine the pedestrian rough edges.

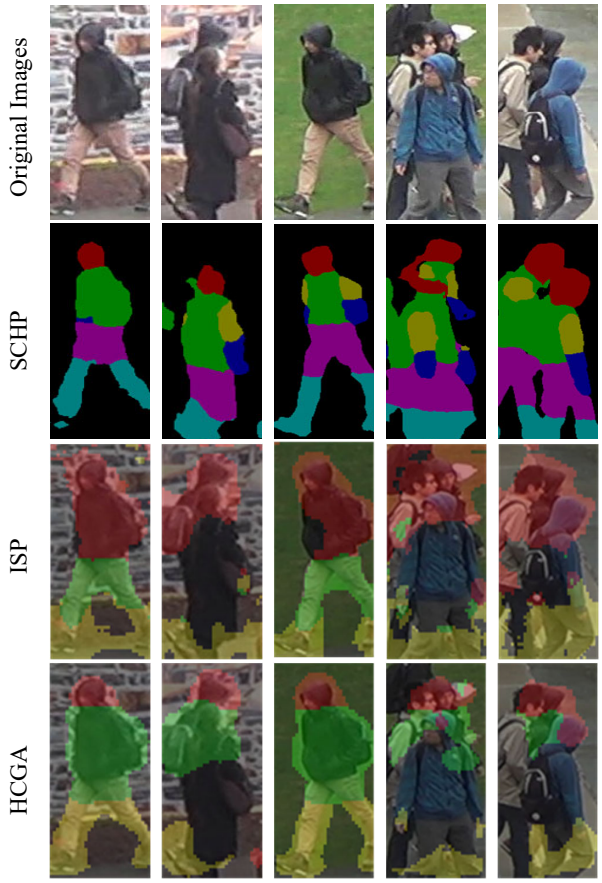

Fig. 9. Comparison of parsing results with existing pixel-level alignment-based methods. All images are the same ID.

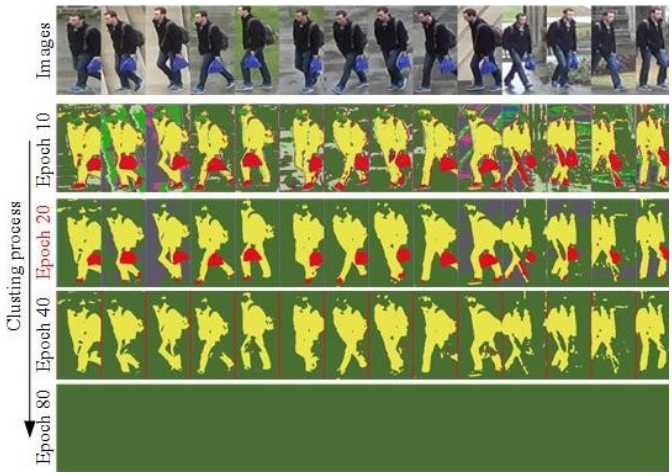

Fig. 10. Visualization of the clustering process of HCNet trained directly on RGB images with the same ID. Different colors indicate different clustering categories, where green and gray are background categories and yellow and red are foreground categories. The training process is represented from top to bottom and is best viewed in color .

B. Human Co-parsing on RGB Images

To illustrate our approach and verify the generality of HCNet, we apply HCNet directly on the RGB image rather than on the feature map extracted by the encoder. The RGB pixel values of a set of images with the same ID are normalized and then fed into the decoder. Notably, the weights λ_L and λ_B in the objective function are different from that of training the HCNet for the person ReID task. As shown in Figure 10, the predicted labels are haphazard at epoch 1. The background and foreground are gradually separated with the network convergence. The best results of segmentation are achieved at epoch 20 when the unique predicted labels converge to a certain number. If the network continues to converge, the objective function is close to 0 and all pixels of images are assigned to one label. This is the reason that we set the minimum number of unique predicted labels in Algorithm 1. Due to the direct use of pixel values as input, the segmentation network may fail when the color and texture of the background and foreground are similar. Therefore, we use the deep network with an encoder-decoder structure. Besides, the proposed method can segment the blue bag in the hand of the pedestrian from the background compared with Human parsing.

VI. Conclusion

In this work, we propose a Human Co-parsing Guided Alignment (HCGA) framework for the task of person ReID. We design local spatial consistency, semantic consistency, and background group losses to weakly supervise the human co-parsing network. The parsing result generated by HCNet guides PRNet to be aligned at the pixel level. PRNet uses a guided alignment module to reduce the uncertainty of segmentation prediction. During inference, only PRNet is used to obtain foreground features and human parts features for matching. Experimental results show that the proposed framework is effective for both holistic and occlusion person ReID problems. Moreover, the visualization results demonstrate that weakly supervised human co-parsing has great potential for occluded person ReID.

Acknowledgment

The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions.

References

[1] G. Shaogang, C. Marco, Y. Shuicheng, and C. L. Chen, Eds., *Person Re-Identification*, ser. Advances in Computer Vision and Pattern Recognition, 2014, pp. 1–445.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and H. S. C. H., "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2021.

[3] S. Yifan, Z. Liang, Y. Yi, T. Qi, and W. Shengjin, "Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of European Conference of Computer Vision*, 2018, pp. 501–18.

[4] L. Zhao, X. Li, Y. Zhuang, J. Wang, and Ieee, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 3239–3248.

[5] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, Conference Proceedings, pp. 7398–7407.

[6] G. Shang, W. Jingya, L. Huchuan, and L. Zimo, "Pose-guided visible part matching for occluded person reid," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 741–9.

[7] W. Guan'an, Y. Shuo, L. Huanyu, W. Zhicheng, Y. Yang, W. Shuliang, Y. Gang, Z. Erjin, and S. Jian, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6448–57.

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

[9] G. Lishuai, Z. Hua, G. Zan, G. Weili, C. Zhiyong, and W. Meng, "Texture semantically aligned with visibility-aware for partial person re-identification," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3771–9.

[10] H. Huang, X. Chen, and K. Huang, "Human parsing based alignment with multi-task learning for occluded person re-identification," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2020, pp. 1–6.

[11] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[12] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[13] Z. Kuan, G. Haiyun, L. Zhiwei, T. Ming, and W. Jinqiao, "Identity-guided human semantic parsing for person re-identification," in *Proceedings of European Conference of Computer Vision*, 2020, pp. 346–63.

[14] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018, pp. 7297–7306.

[15] W.-C. Hung, V. Jampani, S. Liu, P. Molchanov, M.-H. Yang, and J. Kautz, "Scops: Self-supervised co-part segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, 2019, pp. 869–878.

[16] M. Jiaxu, W. Yu, L. Ping, D. Yuhang, and Y. Yi, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 542–51.

[17] Z. Jiaxuan, C. Zeyu, L. Jianhuang, and W. Guangcong, "Occluded person re-identification," ser. Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2018, p. 6 pp.

[18] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[20] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1318–1327.

[21] Y. Hantao, Z. Shiliang, H. Richang, Z. Yongdong, X. Changsheng, and T. Qi, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–71, 2019.

[22] Z. Feng, D. Cheng, S. Xing, J. Xinyang, G. Xiaowei, Y. Zongqiao, H. Feiyue, and J. Rongrong, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15-20 June 2019*, 2019, pp. 8506–14.

[23] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, 2019, pp. 371–381.

[24] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2285–2294.

[25] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5728–5737.

[26] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 5, pp. 1592–1601, 2019.

[27] W. Wang, F. Zhao, S. Liao, and L. Shao, "Attentive waveblock: complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond," *IEEE Transactions on Image Processing*, vol. 31, pp. 1532–1544, 2022.

[28] M. Ye, C. Chen, J. Shen, and L. Shao, "Dynamic tri-level relation mining with attentive graph for visible infrared re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 386–398, 2022.

[29] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. Hoi, "Collaborative refining for person re-identification with label noise," *IEEE Transactions on Image Processing*, vol. 31, pp. 379–391, 2021.

[30] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *IEEE/CVF International Conference on Computer Vision, ICCV*. IEEE, 2021, pp. 14 993–15 002.

[31] K. Zhu, H. Guo, S. Zhang, Y. Wang, G. Huang, H. Qiao, J. Liu, J. Wang, and M. Tang, "Aaformer: Auto-aligned transformer for person re-identification," *CoRR*, vol. abs/2104.00921, 2021.

[32] C. Zhao, X. Lv, S. Dou, S. Zhang, J. Wu, and L. Wang, "Incremental generative occlusion adversarial suppression network for person reid," *IEEE Transactions on Image Processing*, vol. 30, pp. 4212–4224, 2021.

[33] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," ser. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2021.

[34] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Transactions on Multimedia*, 2022.

[35] X. Ji, A. Vedaldi, and J. Henriques, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9864–9873.

[36] W. Kim, A. Kanezaki, and M. Tanaka, "Unsupervised learning of image segmentation based on differentiable feature clustering," *IEEE Transactions on Image Processing*, vol. 29, pp. 8055–8068, 2020.

[37] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2020.

[38] Y. Li, P. Hu, Z. Liu, D. Peng, J. Tianyi Zhou, and X. Peng, "Contrastive clustering," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2021.

[39] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3182–3189.

[40] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, 2019, pp. 8838–8847.

[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of International Conference on Machine Learning*, vol. 1, 2015, pp. 448–456.

[42] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[44] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, vol. abs/1703.07737, 2017.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, and Ieee, "Rethinking the inception architecture for computer vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[46] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 13 001–13 008.

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

[48] L. He, Z. Sun, Y. Zhu, and Y. Wang, "Recognizing partial biometric patterns," *CoRR*, vol. abs/1810.07399, 2018.

[49] M. Jia, X. Cheng, Y. Zhai, S. Lu, S. Ma, Y. Tian, and J. Zhang, "Matching on sets: Conquer occluded person re-identification without alignment," in *Proceedings of Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021, pp. 1673–1681.

[50] Y. F. Sun, L. Zheng, W. J. Deng, S. J. Wang, and Ieee, "Svdnet for pedestrian retrieval," ser. Proceedings of IEEE International Conference on Computer Vision, New York, 2017, pp. 3820–3828.

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

[52] Z. Jiaxuan, L. Jianhuang, and C. Peijia, "A novel teacher-student learning framework for occluded person re-identification [arxiv]," *arXiv*, p. 9 pp., 2019.

[53] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, and Acm, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 2018 ACM Multimedia Conference*, 2018, Book, pp. 274–282.

[54] H. Park and B. Ham, "Relation network for person re-identification," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2020, pp. 11 839–11 847.

[55] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of European Conference of Computer Vision*, vol. 11218, 2018, pp. 418–437.

[56] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3641–3650.

[57] H. Lingxiao, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, 2019, pp. 8449–8458.

[58] C. Ding, K. Wang, P. Wang, and D. Tao, "Multi-task learning with coarse priors for robust part-aware person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1474–1488, 2022.

[59] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, G. Wang, and Ieee, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5363–5372.

[60] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3297–3307.

[61] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3183–3192.

[62] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-reid: Searching for a part-aware convnet for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3750–3759.

[63] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3691–3701.

[64] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 667–676.

[65] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8514–8522.

[66] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Understand scene categories by objects: A semantic regularized scene classifier using

convolutional neural networks," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 2318–2325.
[67] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5686–5696.
[68] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[69] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, and Ieee, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, New York, 2016, pp. 770–778.

Shuguang Dou is currently pursuing the Ph.D. degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, X-ray, and person re-identification.

Cairong Zhao is currently a Professor of College of Electronic and Information Engineering at Tongji University. He received a Ph.D. degree from Nanjing University of Science and Technology, an M.S. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences and a B.S. degree from Jilin University, in 2011, 2006 and 2003, respectively. He works on visual and intelligent learning, including computer vision, pattern recognition and visual surveillance. He has published more than 40 papers in TIP, TIFS, TNNLS, TMM, TCSVT, AAAI, ACMMM and other top journals and conferences on computer vision and artificial intelligence.

Xinyang Jiang received his PhD degree in computer science and technology from Zhejiang University in 2017, and was a senior researcher in Tencent Youtu Lab. He is a researcher at Microsoft Research Asia. His main research areas include cross-modal retrieval, computer vision, and pedestrian re-identification. He has published more than ten papers in CVPR, ECCV, AAAI, ACMMM, TIP and other top conferences and journals on computer vision and artificial intelligence. He is a program member of AAAI, CVPR, MM and other conferences, and a reviewer for TCSVT, TIP and other journals.

Shanshan Zhang (Member, IEEE) received the master's degree in signal and information processing from Tongji University in 2011 and the Ph.D. degree in computer science from the University of Bonn in 2015. From January 2015 to December 2016, she was a Postdoctoral Researcher with the Department of Computer Vision and Multimodal Computing, Max Planck Institute for Informatics, Saarbrucken, Germany. Since December 2016, she has been a Professor with the Nanjing University of Science and Technology, China. Her main research interests include computer vision, pattern recognition, and particularly applications for driverless vehicles.

Wei-Shi Zheng received the Ph.D. degree in applied mathematics from Sun Yat-Sen University, Guangzhou, China, in 2008. He is currently a Full Professor with Sun Yat-Sen University. His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. He has authored/coauthored more than 120 papers, including more than 90 publications in main journals (TPAMI, IJCV, TNN/TNNLS, TIP, and PR) and top conferences (ICCV, CVPR, IJCAI, and AAAI). He is an Associate Editor for the Pattern Recognition Journal and Area Chairs of a number of top conferences. He has joined the Microsoft Research Asia Young Faculty Visiting Program. He is a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of Royal Society-Newton Advanced Fellowship of United Kingdom.

Wangmeng Zuo (Senior Member, IEEE) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. From 2004 to 2006, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. From 2009 to 2010, he was a Visiting Professor with Microsoft Research Asia. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He has published over 90 articles in top-tier academic journals and conferences. His current research interests include image enhancement and restoration, visual tracking, and image classification. He has served as a Tutorial Organizer in ECCV 2016, an Associate Editor of the IET Biometrics, and the Guest Editor of Neurocomputing, Pattern Recognition, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.