Contrastive Pedestrian Attentive and Correlation Learning Network for Occluded Person Re-Identification

Liying Gao*, Bingliang Jiao*, Yuzhou Long, Kai Niu, He Huang†, Peng Wang†, Yanning Zhang

***Abstract*—Occluded person Re-identification (ReID) aims to match occluded and holistic pedestrian images across different camera views. This task presents two primary challenges. First, it is crucial to accurately capture pedestrian foregrounds from seriously occluded person images. Second, a noticeable information asymmetry exists between the partial body in occluded images and the complete body in corresponding holistic images, which could cause the ReID model to underestimate their similarities. To address these challenges, we introduce a contrastive pedestrian attentive and correlation learning (CpaCol) model. Within CpaCol, we first design a Contrastive Pedestrian Attention (ContrastAttn) module to capture pedestrian foregrounds from occluded images. In this process, we notice that most existing attention-based methods only supervise the final predictions with identity loss yet neglect its causality with the generated attention maps, which could mislead the model to capture some salient yet pedestrian-irrelevant noises as discriminative clues. To rectify this, we integrate contrastive learning into our ContrastAttn module to guide it to learn the semantic divergence between pedestrian foregrounds and noises, thereby capturing pedestrian foregrounds more accurately. Besides, we propose a correlation learning module, where we tailor an effective dense feature correlation learning tool, 4D convolution, to enable it to adapt to pedestrian images and capture corresponding clues between comparing images. By focusing more on corresponding clues, our model could avoid overemphasizing the inherent information asymmetry between occluded and holistic images, thereby improving re-identification. Empowered by these modules, our CpaCol achieves state-of-the-art performance on three relevant ReID settings, *i.e.*, occluded, partial, and holistic ReID. Our code is available in https://github.com/nwpugaoliying/CpaCol.**

I. Introduction

Person re-identification (ReID) aims to match the same pedestrian from non-overlapping camera views, which has been widely used in intelligent surveillance and tracking [5], [6]. Over the years, with the rapid development of deep learning and the availability of large-scale ReID datasets, numerous works [3], [7]–[12] have been proposed to address person ReID, achieving remarkable performance. However, most of these methods assume that a complete and unobstructed view of the human body is visible in most images. Unfortunately, in real-world scenarios, this assumption is often violated due to various occlusions, such as trees, cars, and walls. Thus, recently, many studies [13], [14] have been proposed to tackle

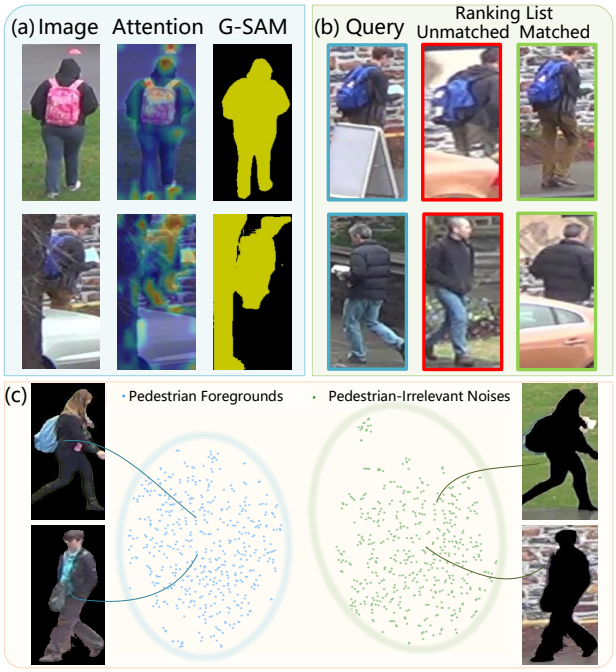

Fig. 1. (a) Existing methods fall short in capturing pedestrian foregrounds on the challenging occluded person images. For example, attention mechanisms may focus on discriminative occlusions and background noises. For pre-trained models like Grounded-SAM [1], [2] (G-SAM), it works well on holistic person images while performing poorly in seriously occluded scenarios. (b) The information asymmetry between the partial body in occluded images and the complete body in holistic images causes their inherent semantic divergence, which could make ReID models underestimate their similarities, thus resulting in biased predictions. Specifically, as shown in the second row, the model aligns the holistic query with the unmatched holistic image instead of the matched occluded one. (c) T-SNE results of pedestrian foreground features and pedestrian-irrelevant noise features. In this part, we first select 500 samples from DukeMTMC-reID [3] dataset and use G-SAM to segment pedestrian foregrounds and pedestrian-irrelevant noises, then employ pre-trained ResNet-50 [4] to extract their features for t-SNE visualization. The results reveal two crucial insights. First, a clear boundary exists between the pedestrian-relevant and pedestrian-irrelevant clusters, indicating the large semantic divergence between features of pedestrian foregrounds and pedestrian-irrelevant noises. Second, significant semantic consistency exists within the foreground (and pedestrian-irrelevant noise) features, leading them to form a tight cluster.

The first two authors contribute equally. Liying Gao is with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: gaoliying@mail.nwpu.edu.cn). Bingliang Jiao, Yuzhou Long, Peng Wang and Yanning Zhang are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China, and Ningbo Institute, Northwestern Polytechnical University, Ningbo 315000, China (email: bingliang.jiao,lyz_@mail.nwpu.edu.cn; peng.wang, ynzhang@nwpu.edu.cn). Kai Niu is with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China, and Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518063, China (email: kai.niu@nwpu.edu.cn). He Huang is with the School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China (email: huanghe1984@nwpu.edu.cn).

†He Huang and Peng Wang are the corresponding authors.

the occluded person ReID task.

In occluded person ReID, we face two major challenges. The first one is to accurately capture pedestrian foregrounds from seriously occluded images for re-identification. To achieve this, many existing methods employ pre-trained pose estimator [14]–[16] or human parsing model [17] to locate the unoccluded human parts. However, these models may underperform in heavily occluded scenarios, as shown in the occluded image of Fig. 1 (a), due to the domain gap between pre-training datasets and target occluded images. Other methods [18]–[21] suggest using attention mechanisms to capture salient pedestrian foregrounds. However, these attention-based methods only supervise the final predictions using identity-level losses, *e.g.*, identity classification loss, neglecting the causality between the predictions and the generated attention maps. As shown in the "Attention" column in Fig. 1 (a), this oversight could potentially mislead the ReID model into treating discriminative occlusions and backgrounds as important clues for re-identification. Hence, we need to explore an effective approach to guide the attention module to accurately disentangle pedestrian foregrounds from irrelevant noises.

The second challenge is the significant information asymmetry existing between occluded and holistic images. Specifically, only partial pedestrian body is visible in occluded images, which inherently causes semantic divergence with the complete body in corresponding holistic images. This divergence can lead the ReID model to underestimate the similarities between occluded and holistic images, resulting in biased predictions. As shown in the first case of Fig. 1 (b), this semantic divergence misleads the ReID model to retrieve an unmatched occluded image for an occluded query instead of its matched holistic image. This mismatch could affect the re-identification, and we thus need to alleviate the interference caused by this inherent information asymmetry.

In this work, we aim to tackle the aforementioned two challenges to facilitate occluded person ReID. For the first challenge, *i.e.*, accurately capturing pedestrian foregrounds, our solution is to equip the attention module with effective guidance to distinguish pedestrian foregrounds from irrelevant noises. Through our investigation, we have made two pivotal observations depicted in Fig. 1 (c): 1) there exists a large semantic divergence between pedestrian foregrounds and pedestrian-irrelevant noises, and 2) significant semantic consistency can be found among pedestrian foregrounds (and pedestrian-irrelevant noises) across varying images. Inspired by these insights, we consider guiding the attention module to learn such semantic divergence and consistency by aligning its focus with these observations, thereby enabling it to capture foregrounds effectively. To accomplish this, we partition pedestrian images into two components: 1) regions focused on by the attention module, which we hope pertain to pedestrian foregrounds, and 2) regions not focused on by the attention module, which should correspond to pedestrian-irrelevant noises, *i.e.*, occlusions and backgrounds. Then, we propose a contrastive learning strategy to align these two components with the observations, in which we enforce the divergence between distinct components as well as the consistency within identical components. This could implicitly encourage our model to accurately capture pedestrian foregrounds.

To address the second challenge, *i.e.*, information asymmetry, we propose to guide the ReID model to capture corresponding pedestrian clues between comparing images. By paying more attention to the fine-grained correspondences, the ReID model could prevent overemphasizing the inherent overall divergence between occluded and holistic images, thus avoiding biased predictions. Notably, this step is non-trivial because we could neither access any pedestrian semantic annotations (*e.g.*, key points, human parsing masks) nor utilize pre-trained semantic models to explicitly locate pedestrian clues within occluded images due to the subpar performance of these models in such contexts. Therefore, in this work, we suggest employing dense local feature similarity measurement, which could aid the ReID model in identifying corresponding pedestrian clues without necessitating additional annotations or extra semantic models.

Practically, in this work, we introduce a Contrastive Pedestrian Attentive and Correlation Learning (CpaCol) model for occluded person ReID, which consists of two essential modules, namely the Contrastive Pedestrian Attention (ContrastAttn) module and the Correlation Learning (CorreL) module. Within ContrastAttn, we leverage contrastive learning to enforce the semantic divergence between focused and non-focused regions of the attention module, while ensuring consistency among focused (and non-focused) regions across different images. This could implicitly encourage the attention module to discern the divergence between pedestrian foregrounds and irrelevant noises, which, however, has a risk of overstressing the semantic divergence, potentially leading the attention module to focus solely on a specific pedestrian part, such as the head. To mitigate this, we divide pedestrian images into multiple horizontal stripes and apply contrastive loss across them. This could enforce the attention module to learn the consistency between pedestrian parts within different stripes, ensuring a comprehensive capture of pedestrian foregrounds. Besides, within the CorreL module, we utilize a well-evaluated dense feature correlation learning tool, 4D convolution [22], [23], to capture corresponding clues between comparing images. Given that corresponding pedestrian clues often appear in nearby positions in comparing images, we introduce a position embedding matrix into the 4D convolution module to enable it could leverage the spatial information of comparing pedestrian clues, thereby capturing correspondences more accurately. By guiding our model to focus on these identified correspondences between comparing images, we could alleviate the impact of inherent semantic divergence between occluded and holistic images.

The main contributions of this paper can be summarised as follows:

- We propose a contrastive pedestrian attentive and correlation learning (CpaCol) network for occluded person ReID. This network can effectively capture pedestrian foregrounds and mitigate the influence of information asymmetry between occluded and holistic images.
- A contrastive pedestrian attention module is proposed, in which we integrate contrastive learning to concentrate the attention module on the distinction between pedes-

trian foregrounds and noises, consequently enabling it to capture the foregrounds effectively.
- A correlation learning module is designed to help our model capture and focus on the correspondences between comparing images. This could prevent our model from making biased predictions influenced by the inherent semantic divergence between occluded and holistic images.
- Extensive experiments verify the effectiveness of our proposed modules. Besides, our CpaCol achieves promising performance on several occluded, partial, and holistic person ReID datasets.

II. Related Work

In this section, we briefly review the relevant works that are related to person re-identification, occluded person re-identification, contrastive learning, and 4D convolution, respectively.

A. Person Re-identification

Person re-identification (ReID) aims to match the same pedestrians captured from non-overlapping camera views. In recent years, abundant works [3], [7], [8], [24]–[27] have been proposed to address the person ReID task due to its potential application in intelligent surveillance and tracking. The existing person ReID methods can be briefly classified into two categories: global feature learning methods and local feature learning methods.

Global feature learning methods strive to develop robust feature extractors or leverage additional semantic guidance to learn effective global-level representations of pedestrian images. For example, Zhou *et al*. [28] introduced a dynamic framework OSNet, which can adaptively extract and integrate multi-scale pedestrian features for identification. Besides, inspired by ViT [29], He *et al*. [30] developed a transformer-based method named TransReID. This method partitions an image into multiple small patches and learns long-range dependencies between these local patches to enhance the robustness of the final extracted features. In addition to the design of effective backbone models, some studies [31], [32] are devoted to using explicit semantic guidance to help ReID models extract noise-free person features. For example, Guo *et al*. [31] proposed P^2Net that employs a human parsing model to produce human masks, which are utilized to extract the holistic features of pedestrian foregrounds.

Local feature learning methods focus on extracting local features of discriminative pedestrian clues for identification. These methods typically fall into two categories: part-based methods and attention-based methods. Part-based methods [30], [33]–[36] generally divide an entire person image into several local regions based on human body structure and then extract local features from these individual regions. For instance, Sun *et al*. [34] horizontally divided the whole image feature into several stripes and then extracted fine-grained local features from each stripe for re-identification. Regarding attention-based methods [37]–[39], they always leverage attention mechanisms to capture discriminative pedestrian clues for re-identification. For instance, Yang *et al*. [38] introduced a class activation maps augmentation (CAMA) model, employing CAMs [40] in different branches to identify various discriminative pedestrian clues. Furthermore, Zhu *et al*. [39] proposed a dual cross-attention learning (DCAL) model, which incorporates a global-local cross-attention module to facilitate interactions between the global image and local regions, so as to capture discriminative pedestrian regions. Additionally, it includes a pairwise cross-attention module to establish interactions between comparing images, promoting the recognition of subtle visual differences.

Although these methods have demonstrated significant advancements in holistic person ReID, they exclusively address scenarios where all body parts are clearly visible, neglecting the potential interference caused by occlusions. Typically, occlusions could introduce significant distortions to pedestrian images, which may lead these methods less effective in capturing pedestrian information for identification. Consequently, in this work, we aim to develop a robust ReID model to handle occluded images effectively.

B. Occluded Person Re-identification

The occluded person ReID task aims to develop robust ReID models to overcome the interference introduced by occlusions and match occluded and holistic person images across different camera views. Typically, existing methods could be primarily divided into two categories: extra semantic-based methods [14]–[17], [41] and attention-based methods [18]–[21], [42], [43].

Among the **extra semantic-based methods**, PVPM [15] adopts a pose estimator, Openpose [44], to generate pose information and subsequently extract pedestrian foreground features with a pose-guided attention module. Additionally, HPNet [17] employs a human parsing model to capture various human parts and then extracts their features. In addition, PFD [41] utilizes a pre-trained human pose estimator to locate key human joints on occluded person images, which subsequently guides the ReID model to disentangle pedestrian features from irrelevant ones. However, due to the domain gap between the pre-training data of these extra semantic models and target occluded images, the employed models could fall short in seriously occluded images (as shown in Fig. 1 (a)), thereby affecting re-identification.

To overcome this limitation caused by the generalization gap, many **attention-based methods** [18]–[21] are proposed to replace pre-trained semantic models with trainable attention mechanisms to capture discriminative pedestrian clues. For example, Part-Aware Transformer (PAT) [18] employs a group of learnable prototypes as queries to capture diverse and discriminative pedestrian clues in occluded images. Besides, Occlusion-Aware Mask Network (OAMN) [19] utilizes an occlusion augmentation scheme to produce diverse occluded images and an attention-guided mask module to capture non-occluded body parts. This approach could enhance the robustness of the ReID model toward various occlusions. In addition, Xu *et al*. [20] proposed to employ a fixed partition strategy to obtain local features and then utilize channel-wise attention operations to disentangle identity-relevant features from these

extracted local features. However, existing attention-based methods merely supervise the final prediction of ReID models using identity-level losses, often overlooking the causality between the predictions and the generated attention maps, which could mislead ReID models to consider discriminative occlusions and backgrounds as crucial clues for identification. To resolve this, in this work, we propose a contrastive learning strategy that focuses the attention module on the distinction between pedestrian foregrounds and noises, consequently enabling it to capture the foregrounds effectively. Besides, in this work, we further design a correlation learning module to alleviate the impact caused by information asymmetry between occluded and holistic images, which is also effective for occluded ReID.

C. Contrastive Learning

Contrastive learning [45]–[47] is an effective metric learning algorithm that could guide deep models to investigate the discrepancies and consistencies among comparing inputs. For instance, He *et al*. [48] proposed utilizing contrastive learning to force deep models to capture the consistency between original images and their augmented versions. This method aids deep models in identifying prominent visual clues and improving their representational capability. Given the efficacy of contrastive learning, it has been widely employed in many ReID works [49], [50]. For example, Chen *et al*. [50] introduced a joint generative and contrastive learning model that uses a GAN model to generate images of the same pedestrian instance from different viewpoints. Then, they treated these images with diverse viewpoints as positive categories and used a contrastive loss to train the ReID model to capture the consistency between them, consequently extracting viewpoint-invariant features. In the field of object part parsing, contrastive learning is also commonly employed to help deep models identify different object parts. For instance, Choudhury *et al*. [51] introduced contrastive learning into the training of the part-aware object parsing model to help the model discern the semantic divergence between different object parts. Inspired by these existing works, we consider employing contrastive learning to assist the ReID model in discovering the semantic divergence between pedestrian foregrounds and pedestrian-irrelevant noises, enabling ReID models to capture pedestrian foregrounds effectively.

D. 4D Convolution

4D convolution [52] is a well-evaluated tool to learn fine-grained correspondences between comparing image pairs. Typically, 4D convolution measures the correlation between two pixels of comparing images by assessing the dense similarity of the pixels in their vicinity. Owing to its effectiveness, 4D convolution has been utilized extensively in many dense correspondence prediction works [52]–[54]. Recently, Min *et al*. [22] proposed a revised version named center-pivot 4D convolution, which replaces dense similarity with center-surrounding similarity to reduce computational consumption. Moreover, Lee *et al*. [23] discovered that incorporating a 4D convolution module could aid retrieval models in capturing fine-grained correspondences between comparing image pairs, consequently enhancing image matching. Drawing inspiration from these works, in this paper, we consider employing 4D convolution to help our model capture and focus on the correspondences between occluded and holistic images, thus alleviating the impact caused by information asymmetric.

III. Approach

In this section, we elaborate on the proposed Contrastive Pedestrian Attentive and Correlation Learning (CpaCol) network. First, we give an introduction to the overall structure. After that, two major components of the CpaCol, *i.e*., the contrastive pedestrian attention module and correlation learning module, are illustrated in order. Finally, some details of training and inference are given.

A. Overall Structure

In this work, we introduce a Contrastive Pedestrian Attentive and Correlation Learning (CpaCol) model for the occluded person ReID task. This task presents two significant challenges, *i.e*., capturing pedestrian foregrounds from seriously occluded person images and mitigating the impact caused by inherent information asymmetry between occluded and holistic images. Our CpaCol model is equipped with two modules, namely, the contrastive pedestrian attention (ContrastAttn) module and the correlation learning (CorreL) module, specifically designed to address the above two challenges.

Fig. 2 provides an overview of our CpaCol model, from which we can find that our CpaCol is trained in a two-stage manner. During training stage I, we optimize the pedestrian feature extraction model, which is a Vision Transformer (ViT) [29] backbone augmented with our ContrastAttn layers. Regarding architecture, three ContrastAttn layers are inserted into the 3rd, 5th, and 7th transformer encoder layers of the ViT. These ContrastAttn layers are used to capture pedestrian foregrounds from occluded images. To prompt our ContrastAttn module to distinguish pedestrian foregrounds from pedestrian-irrelevant noises, we employ a contrastive learning strategy, whose details will be illustrated in Section III-B. After training, this pedestrian feature extraction model would be used to extract pedestrian foreground features to measure the global-level similarities between comparing images.

In addition to the ContrastAttn module, in our CpaCol, we also propose a CorreL module, which could aid our model in identifying and focusing on fine-grained corresponding clues between comparing images, thereby avoiding the ReID model from overemphasizing the global-level information asymmetry between the occluded and holistic images. Specifically, in training stage II, we freeze the pedestrian feature extraction model and utilize the image features extracted from three ContrastAttn layers to train the CorreL module. In this module, we deploy a robust dense feature correlation learning tool, 4D convolution [22], [23], to identify corresponding clues between comparing images. The details of the CorreL module will be given in Section III-C. Given these identified corresponding clues, we then amalgamate the local similarities of these clues into the global-level similarities of comparing images. This

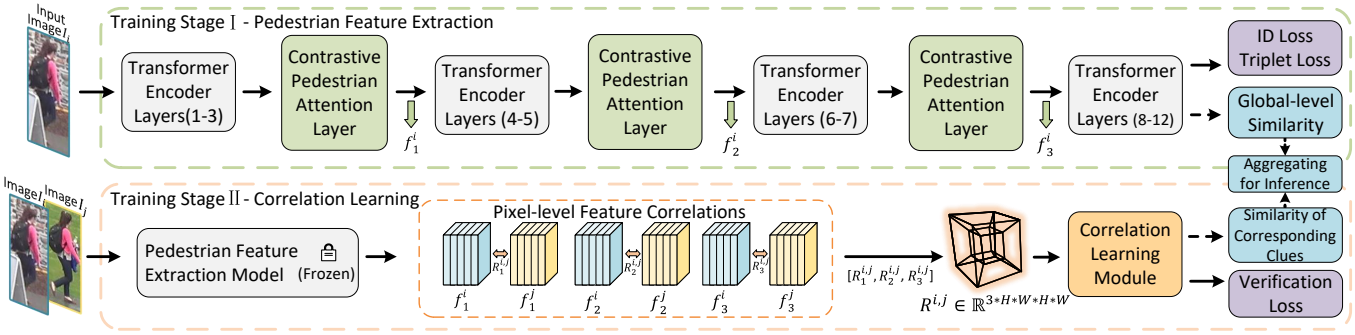

Fig. 2. Illustration of the overall structure of our proposed CpaCol model. The CpaCol model is trained in a two-stage manner, in which we train the pedestrian feature extraction model and correlation learning module, respectively. The image features f_1^i, f_2^i, and f_3^i are extracted from three contrastive pedestrian attention layers and used for fine-grained correlation learning. Note that during training stage II, the pedestrian feature extraction model trained in stage I is frozen. The dotted arrows denote the process for inference. $[,]$ means the concatenation operation.

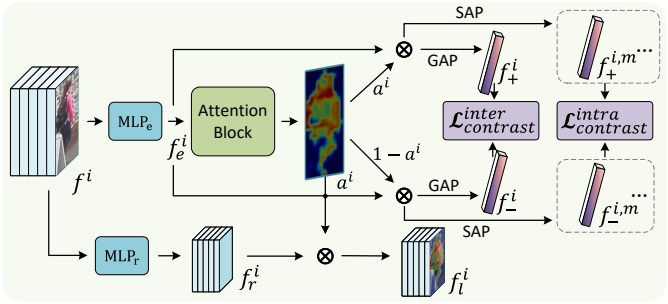

Fig. 3. Illustration of the contrastive pedestrian attention layer. The "$\mathrm{MLP_e}$" and "$\mathrm{MLP_r}$" are two multi-layer perceptron modules, $\otimes$ means a multiplication operation between image features and the attention maps. "GAP" and "SAP" represent a spatial global average pooling and stripe average pooling operation to obtain global-level image features and stripe-level features, respectively. $\mathcal{L}_{contrast}^{inter}$ and $\mathcal{L}_{contrast}^{intra}$ denote the inter-image contrastive loss and intra-image contrastive losses.

approach could guide our model to pay more attention to the corresponding clues between comparing images, thereby mitigating the impact of information asymmetry.

B. Contrastive Pedestrian Attention Module

In occluded person ReID, attention-based methods [18]–[21] are frequently utilized to identify discriminative pedestrian clues. However, most of these methods only supervise the final prediction using identity loss, largely neglecting the causality between the prediction and the generated attention map. This oversight may mislead the attention module to treat some discriminative occlusions and backgrounds as crucial clues. To mitigate this, we introduce a Contrastive Pedestrian Attention (ContrastAttn) module, in which we employ a contrastive learning strategy to help the attention module discern the divergence between pedestrian foregrounds and irrelevant noises, thereby enabling it to capture pedestrian foregrounds more accurately.

In this module, our main objective is to formulate an effective strategy that could guide the attention module to precisely capture pedestrian foregrounds. During this process, we find two crucial observations, as shown in Fig. 1 (c): 1) There is a significant semantic divergence between pedestrian foregrounds and pedestrian-irrelevant noises. 2) Substantial semantic consistency exists within pedestrian foregrounds (and pedestrian-irrelevant noises) across various images. Given these observations, we consider guiding the attention module to learn such divergence and consistency by aligning its focus accordingly, hence enabling it to accurately distinguish the pedestrian foregrounds from irrelevant noises. To implement this, in this module, we propose a contrastive learning strategy to enforce the attention module to amplify the semantic divergence between its focused and non-focused contents, while enlarging the semantic consistency among focused (non-focused) content across varying images. In this step, we notice that semantic divergence exists not only between pedestrian foregrounds and noises but also among various pedestrian body parts. Overemphasizing the semantic divergence could mislead the attention module to focus solely on some specific pedestrian body parts rather than the entire foreground. To address this concern, we follow [34] to partition pedestrian images into K horizontal stripes and further apply contrastive learning to enforce the attention module to capture semantically consistent pedestrian contents within each stripe, thereby equipping it to comprehensively capture the pedestrian foregrounds.

The illustration of the ContrastAttn layer is exhibited in Fig. 3. Practically, given the image feature $f^i \in \mathbb{R}^{HW \times C}$ of image I_i obtained from a transformer encoder layer, where HW is the number of image tokens, and C is the channel dimension of ViT ($C = 768$), we first apply a spatial attention operation over it to generate the attention map. In this process, the image feature f^i is first embedded by a multi-layer perceptron module $\mathrm{MLP_e}$, which consists of three fully-connected layers inserted with a layer normalization operation and a GELU activation function [55]. The embedded feature is then reshaped into $f_e^i \in \mathbb{R}^{C \times H \times W}$, which is subsequently sent to an attention block to generate its attention map $a^i \in \mathbb{R}^{1 \times H \times W}$. This step can be presented as,

$$a^i = \mathrm{Sigmoid}(\mathrm{BN}(\mathrm{Conv}(f_e^i))), \tag{1}$$

where $\mathrm{Conv}(\cdot)$, $\mathrm{BN}(\cdot)$, and $\mathrm{Sigmoid}(\cdot)$ denote a convolutional

layer with an output channel of 1, batch normalization layer, and Sigmoid activation function, respectively. Then we can use the attention map to separate the pedestrian image into two components: 1) the focused regions by the ContrastAttn layer a^i, which should be pedestrian foregrounds, and 2) the non-focused regions $(\mathbf{1} - a^i)$, which we hope refers to pedestrian-irrelevant noises, *i.e.*, occlusions and backgrounds. By applying attended pooling over input features with a^i and $(\mathbf{1} - a^i)$, we can get the features of focused component f_+^i and features of non-focused component f_-^i, respectively. This process could be written as,

$$\begin{aligned} f_+^i &= \mathrm{AvgPool}(f_e^i \otimes a^i), \\ f_-^i &= \mathrm{AvgPool}(f_e^i \otimes (\mathbf{1} - a^i)), \end{aligned} \tag{2}$$

where $\otimes$ and $\mathrm{AvgPool}(\cdot)$ respectively denote a multiplication operation and a spatial average pooling layer.

After that, we apply an inter-image contrastive loss $\mathcal{L}_{contrast}^{inter}$ over the features of these two components to align them with the observations mentioned above. Specifically, the inter-image contrastive loss $\mathcal{L}_{contrast}^{inter}$ consists of two parts, namely, the negative contrastive loss to enlarge the semantic divergence between different components and the positive contrastive loss to enforce the semantic consistency of the same components. Specifically, the negative contrastive loss can be formulated as,

$$\mathcal{L}_{neg}^{inter} = -\sum_{i=1}^{N}\sum_{j=1}^{N} log\left(1 - \left[sim(f_+^i, f_-^j)\right]_+\right), \tag{3}$$

where N is the batch size and $sim(,)$ is the cosine similarity between a pair of feature representations; $[\cdot]_+$ denotes the ReLU function. Then, we apply a positive contrastive loss between focused components (and non-focused components) across images. In this step, we notice that forcing our model to enlarge the semantic consistency between foregrounds of different pedestrian instances might conflict with the primary goal of the ReID task, *i.e.*, learning distinguishable pedestrian features. Hence, we only apply contrastive loss between focused components of pedestrian images with the same identities. The positive contrastive loss could then be represented as,

$$\begin{aligned} \mathcal{L}_{pos\_+}^{inter} &= -\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{1}_{[id(i)=id(j)]} log\left(\left[sim(f_+^i, f_+^j)\right]_+\right), \\ \mathcal{L}_{pos\_-}^{inter} &= -\sum_{i=1}^{N}\sum_{j=1}^{N} log\left(\left[sim(f_-^i, f_-^j)\right]_+\right), \end{aligned} \tag{4}$$

where $\mathbb{1}_{[id(i)=id(j)]}$ is 1 if the identity label of the i-th image is equal to that of the j-th image, otherwise 0. The overall inter-image contrastive loss could be written as,

$$\mathcal{L}_{contrast}^{inter} = \mathcal{L}_{neg}^{inter} + \mathcal{L}_{pos\_+}^{inter} + \mathcal{L}_{pos\_-}^{inter}. \tag{5}$$

In addition to the inter-image contrastive loss, we also apply an intra-image contrastive loss to guide the attention module to concentrate on the semantic consistency among pedestrian body parts. In this step, we partition each pedestrian image into several horizontal stripes as in [34] and conduct contrastive loss among these local stripes. In each stripe, we independently follow Eq. 1 and Eq. 2 to extract features of focused and non-focused components. Given the m-th stripe of image I_i, the local features of focused regions and non-focused regions in the current stripe could be written as $f_+^{i,m}$ and $f_-^{i,m}$, respectively. Then, similar to the inter-image contrastive loss, we apply both positive and negative contrastive losses across different stripes, which could be written as,

$$\begin{aligned} \mathcal{L}_{contrast}^{intra} &= \mathcal{L}_{neg}^{intra} + \mathcal{L}_{pos\_+}^{intra} + \mathcal{L}_{pos\_-}^{intra}, \\ \mathcal{L}_{neg}^{intra} &= -\sum_{m=1}^{K}\sum_{n=1}^{K} log\left(1 - \left[sim(f_+^{i,m}, f_-^{i,n})\right]_+\right), \\ \mathcal{L}_{pos\_+}^{intra} &= -\sum_{m=1}^{K}\sum_{n=1}^{K} log\left(\left[sim(f_+^{i,m}, f_+^{i,n})\right]_+\right), \\ \mathcal{L}_{pos\_-}^{intra} &= -\sum_{m=1}^{K}\sum_{n=1}^{K} log\left(\left[sim(f_-^{i,m}, f_-^{i,n})\right]_+\right), \end{aligned} \tag{6}$$

where K is the number of partitioned stripes in an image. By this means, we could force the attention module to concentrate on the semantic consistency among pedestrian body parts, thus encouraging it to capture complete pedestrian foregrounds.

Finally, we multiply the generated attention map with the original input image feature f^i to obtain pedestrian foreground features. Formally, this step could be written as,

$$f_l^i = \mathrm{MLP_r}(f^i) \otimes \mathrm{Flatten}(a^i), \tag{7}$$

where $\mathrm{MLP_r}$ represents a multi-layer perceptron module used to embed original input features, and $f_l^i \in \mathbb{R}^{HW \times C}$ denotes the pedestrian foreground features extracted by our ContrastAttn layer.

C. Correlation Learning Module

Generally, occluded images only contain partial pedestrian parts, such as the head and upper body, which inevitably causes information asymmetry between occluded images and their holistic counterparts. The semantic divergence brought by this asymmetry could mislead the ReID model to underestimate the similarities between corresponding occluded and holistic images, affecting re-identification. To mitigate this issue, we propose guiding the ReID model to identify and pay more attention to corresponding clues between occluded and holistic images rather than focusing exclusively on their global-level similarities, thereby alleviating the impact caused by the inherent information asymmetry. To facilitate this, in our CpaCol, we design a correlation learning (CorreL) module responsible for identifying corresponding clues. Within this module, a thoroughly evaluated dense feature correlation learning tool, *i.e.*, 4D convolution, is employed to capture corresponding elements between comparing images. Particularly, given that corresponding pedestrian clues often appear in nearby positions within comparing images, in this module, we carefully revise existing 4D convolution to enable it to utilize the spatial information of comparing pedestrian clues, thereby capturing correspondences more accurately.

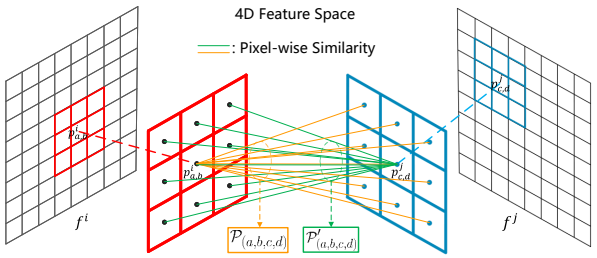

Fig. 4. Illustration of the center-pivot 4D convolution [22], which models the correspondence between the pixels $p^i_{a,b}$ and $p^j_{c,d}$ from two images by considering the similarities between $p^i_{a,b}$ and k-neighborhood pixels of $p^j_{c,d}$ ($\mathcal{P}_{(a,b,c,d)}$) as well as the similarities between $p^j_{c,d}$ and k-neighborhood pixels of $p^i_{a,b}$ ($\mathcal{P}'_{(a,b,c,d)}$) together.

4D Convolution. For easy understanding, in this part, we first briefly illustrate the 4D convolution, a well-evaluated tool to learn fine-grained correspondences between comparing images. Considering the original versions [54], [56] of 4D convolution is computationally intensive, in this work, we employ the center-pivot 4D convolution [22] (CP-4D Conv) to reduce computational consumption. Typically, CP-4D Conv models the correlation between two pixels from distinct images by assessing their center-surrounding similarities. Specifically, given a pair of comparing images I_i and I_j, we could denote the pixel located in coordinate (a, b) in I_i as $p^i_{a,b}$ while representing the pixel located in coordinate (c, d) in I_j as $p^j_{c,d}$. To model the correlation between $p^i_{a,b}$ and $p^j_{c,d}$, CP-4D Conv first calculates the similarities between $p^i_{a,b}$ and the k-neighborhood pixels of $p^j_{c,d}$, which could be denoted as $\mathcal{P}_{(a,b,c,d)} \in \mathbb{R}^{1,\mathrm{k},\mathrm{k}}$, as well as the similarities between $p^j_{c,d}$ and the k-neighborhood pixels of $p^i_{a,b}$, which could be writen as $\mathcal{P}'_{(a,b,c,d)} \in \mathbb{R}^{1,\mathrm{k},\mathrm{k}}$. In Fig. 4, we illustrate the process of computing $\mathcal{P}_{(a,b,c,d)}$ and $\mathcal{P}'_{(a,b,c,d)}$. With $\mathcal{P}_{(a,b,c,d)}$ and $\mathcal{P}'_{(a,b,c,d)}$, CP-4D Conv applies two separate 2D convolution operations over them to model the correlation between $p^i_{a,b}$ and $p^j_{c,d}$. This process can be formally written as,

$$\mathrm{Cor}_{(a,b,c,d)} = \mathrm{Conv2D}(\mathcal{P}_{(a,b,c,d)}) + \mathrm{Conv2D}'(\mathcal{P}'_{(a,b,c,d)}), \tag{8}$$

where $\mathrm{Conv2D}$ and $\mathrm{Conv2D}'$ represent 2D convolutional operations with a kernel size of k, $\mathrm{Cor}_{(a,b,c,d)}$ denotes the correlation between $p^i_{a,b}$ and $p^j_{c,d}$ learned by CP-4D Conv. In this way, CP-4D Conv could take the surrounding information into account to model the correlation between $p^i_{a,b}$ and $p^j_{c,d}$. For more details of CP-4D Conv, please refer to [22].

In our CorreL module, to comprehensively capture the correspondences between I_i and I_j leveraging CP-4D Conv, we need to first calculate the similarities between each pixel pair in comparing images. To achieve this, we utilize features extracted from three incorporated ContrastAttn layers inside the trained pedestrian feature extraction model to construct three dense similarity matrices between I_i and I_j. The rationale for utilizing features extracted from the ContrastAttn layers lies in their ability to eliminate pedestrian-irrelevant noises, which enables our module to accurately capture corresponding pedestrian foreground clues. In this step, the extracted features could be written as, f^i_l and $f^j_l \in \mathbb{R}^{C\times H\times W}$, where $l \in \{1, 2, 3\}$ indicates which ContrastAttn layer the current features are extracted from, as demonstrated in Fig. 2. Formally, the process of similarity measurement could be written as,

$$r^{i,j}_l(a, b, c, d) = \mathrm{ReLU}\left(\frac{\omega_l(f^i_l(a, b)) \cdot \omega_l(f^j_l(c, d))}{||\omega_l(f^i_l(a, b))||\ ||\omega_l(f^j_l(c, d))||}\right), \tag{9}$$

where (a, b) and (c, d) are spatial pixel indices, $\omega_l(\cdot)$ refers to an embedding layer that compresses the channel dimension of f^*_l from C to C_r ($C_r = 256$), thereby reducing computational demands. The variable $r^{i,j}_l(a, b, c, d)$ represents the similarity between the features of pixel $p^i_{(a,b)}$ and the features of pixel $p^j_{(c,d)}$. The function $\mathrm{ReLU}(\cdot)$ is employed to filter out pairs of pixels that exhibit high-level non-correlations. Subsequently, we integrate the similarities between all pixel pairs to construct the dense similarity matrix between f^i_l and f^j_l, which could be represented as $R^{i,j}_l \in \mathbb{R}^{H\times W\times H\times W}$.

Simultaneously, given the spatial correspondences between pedestrian images, *i.e.*, corresponding body parts typically occur in nearby positions in comparing images, we intend to revise the 4D convolution to enable it to exploit the spatial information of comparing pixels, thereby capturing correspondences more accurately. To this end, we propose a trainable position embedding matrix, denoted as $\alpha \in \mathbb{R}^{H,W,H,W}$, where each element indicates the relative spatial correlation between a pair of comparing pixels. Taking the element $\alpha(a, b, c, d)$ as an example, which describes the relative spatial correlation between $p^i_{a,b}$ and $p^j_{c,d}$, its initial value is determined based on the spatial distance between $p^i_{a,b}$ and $p^j_{c,d}$. Specifically, during this step, we divide each image into four horizontal stripes. If the pixels $p^i_{a,b}$ and $p^j_{c,d}$ are located within the same stripes, $\alpha(a, b, c, d)$ is initialized to 1.0. If they are located in adjacent stripes, the initial value is set to 0.8. For a gap of one stripe, the value is initialized to 0.6, while for a gap of two stripes, the initial value is established at 0.4. In this manner, α can describe the relative spatial correlation between each pair of comparing pixels. By multiplying the trainable α by $R^{i,j}_l$, we could allow our model to access and learn to make good use of spatial information for efficient correspondence capture. After applying the position embedding matrix, we then concatenate all three dense similarity matrices together, denoting as $R^{i,j} = [R^{i,j}_1, R^{i,j}_2, R^{i,j}_3] \in \mathbb{R}^{3\times H\times W\times H\times W}$. Notably, the center-surrounding similarities $\mathcal{P}_{()}$ and $\mathcal{P}'_{()}$ between arbitrary pixel pair could be directly extracted from this embedded 4D dense similarity matrix $R^{i,j}$.

After that, we send $R^{i,j}$ into the 4D convolutional layers to measure the correlation between all pixel pairs within the comparing images. In this work, we follow [23] to stack four 4D convolutional blocks. As demonstrated in Fig. 5, we reduce the spatial dimension of $R^{i,j}$ from $H\times W\times H\times W$ to $\frac{H}{8}\times\frac{W}{8}\times\frac{H}{8}\times\frac{W}{8}$ by adjusting the stride of the 2D convolution operation in Eq. 8. Simultaneously, we gradually increase the channel dimension of $R^{i,j}$ from 3 to 128 by adjusting the channel dimension of the convolution kernels in Eq. 8. The final output correlation map between the i-th image and the j-th image is denoted as $R^{i,j}_{out} \in \mathbb{R}^{128\times\frac{H}{8}\times\frac{W}{8}\times\frac{H}{8}\times\frac{W}{8}}$. The elements with

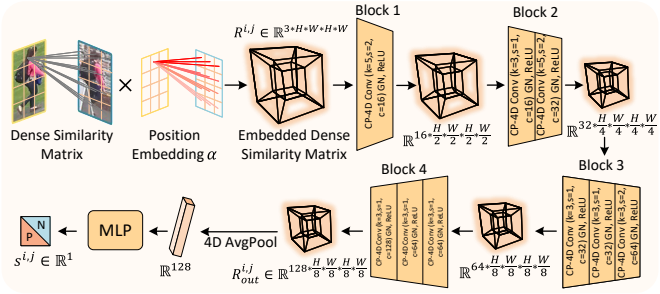

Fig. 5. Illustration of the correlation learning module. It first calculates the embedded similarity matrix $R^{i,j}$ between comparing images, then compress it into an image-level similarity $s^{i,j}$ with four 4D convolutional blocks, a 4D average pooling layer (AvgPool), and a multi-layer perceptron (MLP). In each block, "CP-4D Conv" represents the center-pivot 4D convolution [22], "k", "s", and "c" denote the kernel size, stride, and output channel dimension of the convolutional kernels, and "GN" means a group normalization layer.

high responses could indicate corresponding clues between the comparing images.

To guide our CorreL module to accurately capture correspondences between comparing images, in this work, we follow [23] to train it with a verification loss. Specifically, as shown in Fig. 5, acquiring the $R^{i,j}_{out}$, which models the fine-grained correspondences between comparing images, we first apply a 4D average pooling layer over it. This result is then inputted into an MLP module followed by a Sigmoid function. By this means, the local similarities of all captured corresponding clues are integrated into an image-level similarity $s^{i,j} \in \mathbb{R}^1$. Thereafter, we apply an image verification loss $\mathcal{L}_{ver}$ over it, which could be written as,

$$\mathcal{L}_{ver} = -\left(\mathbb{1}^{i,j} log(s^{i,j}) + (1 - \mathbb{1}^{i,j}) log(1 - s^{i,j})\right), \quad (10)$$

where $\mathbb{1}^{i,j}$ is 1 if i-th image and j-th image belong to the same identities, otherwise 0. Under the guidance of image verification loss, the CorreL module could be encouraged to accurately capture fine-grained correspondences between comparing images. To provide an intuitive illustration of the corresponding clues captured by our CorreL module, we include some visualizations in Fig. 6. During the inference stage, the $s^{i,j}$ is integrated into the global-level similarities between comparing images for prediction. This strategy could help our model to focus more on corresponding clues between occluded and holistic images, thereby avoiding biased predictions.

D. Training and Inference

In training stage I, we aim to train a robust pedestrian feature extraction model, *i.e.*, ViT [29] inserted with our ContrastAttn module. In this step, we follow [30] to extract part-level pedestrian features f_p by applying stripe average pooling [34] over the output image token features of ViT. Then, we apply triplet loss and identity classification loss over both the extracted part-level features and the output features of [cls] token as follows,

$$\begin{aligned}\mathcal{L}_{base} = \mathcal{L}_{id}(\mathrm{FC}_{cls}(f_{cls})) + \frac{1}{K}\sum\nolimits_{k=1}^{K} \mathcal{L}_{id}(\mathrm{FC}^k_p(f^k_p)) \\ + \mathcal{L}_{tri}(f_{cls}) + \frac{1}{K}\sum\nolimits_{k=1}^{K} \mathcal{L}_{tri}(f^k_p),\end{aligned} \quad (11)$$

Fig. 6. Visualization cases of learned fine-grained correlations. We use connecting lines to display the top 10 significant pattern correspondences captured by our CorreL module, which have the highest response values in the learned 4D correspondence map. We can find that our CorreL module could effectively and accurately capture the correspondences between comparing images.

where f_{cls} and f^k_p are the feature of [cls] token and k-th part-level feature; $\mathcal{L}_{id}$ and $\mathcal{L}_{tri}$ denote the identity classification loss and triplet loss, respectively; K is the number of stripes in each person image, which is set to 4, as in [30]; FC_{cls} and FC^k_p are identity classifiers. Besides, we also follow Eq. 5 and Eq. 6 to respectively apply inter-image contrastive loss and intra-image contrastive loss over the ContrastAttn module to encourage it to capture pedestrian foregrounds. Hence, the overall loss function for training stage I is represented as,

$$\mathcal{L}_{stage\ \mathrm{I}} = \mathcal{L}_{base} + \lambda_1 \mathcal{L}^{inter}_{contrast} + \lambda_2 \mathcal{L}^{intra}_{contrast}, \quad (12)$$

where λ_1 and λ_2 denote the loss weight for inter-image contrastive loss $\mathcal{L}^{inter}_{contrast}$ and intra-image contrastive loss $\mathcal{L}^{intra}_{contrast}$, respectively.

In training stage II, we employ the features extracted from the trained ContrastAttn layers to construct similarity matrices, which are then used to train our CorreL module. The CorreL module is trained with a verification loss and utilized to capture corresponding clues between comparing images. The loss function for this stage is denoted as,

$$\mathcal{L}_{stage\ \mathrm{II}} = \mathcal{L}_{ver}. \quad (13)$$

In this stage, for each image, we sample one image with the same identity and another one with a different identity for the verification loss computation.

In the test stage, we amalgamate the global-level image similarity with the similarity of captured corresponding clues together for prediction. Specifically, we first calculate the global-level features between comparing images by concatenating all part-level features f^*_p and features of the class token f_{cls} together, which are then used to measure the global-level image similarity, as in [30]. After that, we amalgamate the integrated local similarity of corresponding clues between comparing images, *i.e.*, the $s^{i,j}$ in Eq. 10, with the global-level similarity together for inference. This could help to focus more on the corresponding clues between comparing images.

IV. Experiment

In this section, we conduct comprehensive experiments to verify the superiority of our proposed CpaCol network for occluded person ReID.

A. Datasets and Evaluation Metrics

We evaluate our model on six ReID datasets for three tasks, including occluded person ReID [13], [14], partial person ReID [57], [58], and holistic person ReID [3], [8].

Occluded-Duke [14] is derived from DukeMTMC-reID [3] dataset for occluded person ReID. It consists of $15,618$ training images, $2,210$ query images, and $17,661$ gallery images. In this dataset, most query images are occluded by a large variety of occlusions, while gallery images contain both holistic and occluded images.

Occluded-REID [13] is an occluded person ReID dataset captured by mobile cameras. It contains $2,000$ images belonging to 200 identities. Each identity has 5 full-body person images and 5 occluded person images with different types of occlusions.

Partial-REID [57] is a partial person ReID dataset collected at a university campus. It contains 600 images of 60 identities, with 5 partial images and 5 holistic images for each identity. The partial images serve as query images, and holistic images are gallery images.

Partial-iLIDS [58] is a partial person ReID dataset derived from iLIDS [59]. It comprises 238 images from 119 identities, with one query image and one gallery image for each identity. This dataset is collected in an airport, where the lower-body parts of pedestrians are often occluded by the luggage.

Market-1501 [8] is a widely-used holistic person ReID dataset. It contains $12,936$ training images of 751 persons, $19,732$ query images, and $3,368$ gallery images of 750 persons captured from 6 cameras.

DukeMTMC-reID [3] is also treated as a holistic person ReID dataset. It contains $16,522$ training images, $2,228$ query images, and $17,661$ gallery images.

Evaluation Metrics. Following the standard evaluation setting, we employ the Cumulative Matching Cure (CMC) and mean Average Precision (mAP) as the evaluation metrics. The values of the CMC curves at rank-1, rank-3, rank-5, and rank-10 are represented as R-1, R-3, R-5, and R-10, respectively.

B. Implementation Details

To make a fair comparison with the state-of-the-art methods [41], [42], we utilize the same backbone, *i.e.*, ViT [29], with a small step sliding-window setting [30], [41]. In our experiments, both training and test images are resized to 256×128. The training images are augmented with random horizontal flipping, padding, random cropping, and random erasing [61]. Following [30], we employ the SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4. In training stage I, the pedestrian feature extraction model, including the ViT backbone and the contrastive pedestrian attention module, is optimized. The learning rate is initialized to 0.08 with cosine learning rate decay. The batch size is set to 64 with 4 images per identity. In training stage II, only the correlation learning module is optimized. We use the same learning rate decay strategy with an initial learning rate of 0.01.

Method	Occluded-Duke		Occluded-REID	
	R-1	mAP	R-1	mAP
PGFA (ICCV 19) [14]	51.4	37.3	-	-
HPNet (ICME 20) [17]	-	-	87.3	77.4
PVPM (CVPR 20) [15]	47.0	37.7	70.4	61.2
HOReID (CVPR 20) [16]	55.1	43.8	80.3	70.2
PFD (AAAI 22)* [41]	69.5	61.8	81.5	83.0
FPR (ICCV 19) [60]	-	-	78.3	68.0
OAMN (ICCV 21) [19]	62.6	46.1	-	-
PAT (CVPR 21) [18]	64.5	53.6	81.6	72.1
TransReID (ICCV 21)* [30]	66.4	59.2	-	-
FED (CVPR 22)* [42]	68.1	56.4	86.3	79.3
FRT (TIP 22) [20]	70.7	61.3	80.4	71.0
PRE-Net (TCSVT 23) [21]	68.3	55.2	-	-
CpaCol (Ours)*	**72.8**	**65.2**	**88.2**	**84.0**

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON OCCLUDED-DUKE AND OCCLUDED-REID. * MEANS THE METHODS USE VIT [29] AS THE BACKBONE MODEL. OUR CPACOL OUTPERFORMS THE COMPARED METHODS BY A LARGE MARGIN ON THESE TWO OCCLUDED REID DATASETS.

C. Comparison with State-of-the-art Methods

Comparison on Occluded Person ReID Datasets. In Table I, we compare the performance of our proposed CpaCol and the previous methods on two occluded person ReID datasets, *i.e.*, Occluded-Duke and Occluded-REID. As shown in Table I, the compared methods are classified into two categories: extra semantic-based methods (the upper block) and attention-based methods (the lower block). Specifically, compared to the best competitor of extra semantic-based methods, *i.e.*, PFD [41], which employs a pre-trained pose estimation model to locate human body parts, our CpaCol outperforms it by a large margin, *i.e.*, 3.3% and 6.7% in R-1 accuracy on Occluded-Duke and Occluded-REID, respectively. The reason for this performance gain could be that the pre-trained pose estimation model employed in PFD does not generalize well to heavily occluded images and hinders the ReID performance, while our ContrastAttn module is trained to adapt to the occluded images and could capture pedestrian contents more accurately. Besides, our CpaCol also outperforms the top-performing attention-based methods FRT [20] by 3.9% of mAP on Occluded-Duke and FED [42] by 4.7% of mAP on Occluded-REID. The reason for these results could be that the compared attention-based methods learn to capture discriminative clues only under the implicit guidance of identity losses, which may mislead them to attend to some discriminative yet pedestrian-irrelevant noises. In contrast, our CpaCol can effectively disentangle pedestrian foregrounds and irrelevant noises under the guidance of our contrastive learning strategy. Meanwhile, our designed CorreL module could also help our model to capture and focus on the fine-grained correspondences between comparing images, consequently enabling it to be robust to information asymmetry between occluded and holistic images. Therefore, our proposed CpaCol outperforms compared methods on both Occluded-Duke and Occluded-REID datasets.

Method	Partial-REID		Partial-iLIDS	
	R-1	R-3	R-1	R-3
PGFA (ICCV 19) [14]	68.0	80.0	69.1	80.9
HPNet (ICME 20) [17]	85.7	-	72.0	-
PVPM (CVPR 20) [15]	78.3	87.7	-	-
HOReID (CVPR 20) [16]	85.3	91.0	72.6	86.4
FPR (ICCV 19) [60]	81.0	-	-	-
SORN (TCSVT 20) [62]	76.7	84.3	79.8	86.6
OAMN (ICCV 21) [19]	86.0	-	77.3	-
PRE-Net (TCSVT 23) [21]	86.0	91.3	78.2	87.4
CpaCol (Ours)	**86.3**	**94.0**	**85.9**	**91.7**

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE PARTIAL-REID AND PARTIAL-ILIDS DATASETS. OUR CPACOL OUTPERFORMS THE COMPARED METHODS ON THESE TWO PARTIAL PERSON REID DATSETS.

Method	Market-1501		DukeMTMC-reID	
	R-1	mAP	R-1	mAP
PCB (ECCV 18) [34]	92.3	77.4	81.8	66.1
DATRL (TCSVT 19) [35]	94.4	81.5	86.3	72.9
BOT (CVPRW 19) [33]	94.1	85.7	86.4	76.4
CBDB (TCSVT 21) [36]	94.4	85.0	87.7	74.3
PGFA (ICCV 19) [14]	91.2	76.8	82.6	65.5
P^2Net (ICCV 19) [31]	95.2	85.6	86.5	73.1
HOReID (CVPR 20) [16]	94.2	84.9	86.9	75.6
FRT (TIP 22) [20]	**95.5**	88.1	90.5	81.7
IANet (CVPR 19) [37]	94.4	83.1	87.1	73.4
CAMA (CVPR 19) [38]	94.7	84.5	85.8	72.9
FPR (ICCV 19) [60]	95.4	86.6	88.6	78.4
PAT (CVPR 21) [18]	95.4	88.0	88.8	78.2
TransReID (ICCV 21) [30]	95.2	88.9	90.7	82.0
FED (CVPR 22) [42]	95.0	86.3	89.4	78.0
PRE-Net (TCSVT 23) [21]	95.3	86.5	89.3	77.8
CpaCol (Ours)	95.2	**89.8**	**91.5**	**85.2**

TABLE III
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON MARKET-1501 AND DUKEMTMC-REID. OUR METHOD ACHIEVES PROMISING PERFORMANCE ON THESE TWO HOLISTIC PERSON REID DATASETS.

Comparison on Partial Person ReID Datasets. To further evaluate our proposed CpaCol, we also conduct experiments on two partial person ReID datasets, *i.e.*, Partial-REID and Partial-iLIDS. In this part, we follow existing works to train our CpaCol model on the training set of Market-1501 [8] and subsequently evaluate it on the above two partial ReID datasets. Table II exhibits the experimental results of our CpaCol and the compared methods. Here, the compared methods could also be classified into extra semantic-based methods (the first block) and attention-based methods (the second block). From the results in Table II, we can find that our CpaCol also outperforms all compared methods on the Partial-REID and Partial-iLIDS datasets. Among the compared methods, the top-performing competitor, PRE-Net [21], focuses on identifying pedestrian foregrounds via attention mechanism but fails to account for the information asymmetry between partial and holistic images. In contrast, our CpaCol model could both mitigate the impact of information asymmetry between partial and holistic images by utilizing the CorreL module and identify pedestrian foregrounds through the ContrastAttn module, thus achieving superior performance on the partial person ReID datasets.

Comparison on Holistic Person ReID Datasets. Given that the ultimate goal of mitigating interference of occlusions is to enhance performance in the ReID task, it is thus also essential to evaluate our algorithm's effectiveness on general holistic person ReID datasets with limited occlusions. Therefore, we further compare our CpaCol model with existing methods on two holistic person ReID datasets, namely, Market-1501 and DukeMTMC-reID. The experimental results are given in Table III. The compared methods could be briefly divided into three categories: part-based methods (the first block), extra semantic-based methods (the second block), and attention-based methods (the third block). From the results, we can find that our CpaCol still achieves promising performance on holistic person ReID datasets. Although our CpaCol's R-1 accuracy on Market-1501 is slightly lower (about 0.3%) than the top-performing model FRT [20], it achieves significantly better mAP performance on both datasets, *i.e.*, 2.6% on average. This promising performance could indicate that our CpaCol is also able to address the interference caused by occlusions in the holistic person ReID datasets.

ContrastAttn		CorreL		O-Duke		O-REID	
Inter.	Intra.	4D Convs	PosEmb.	R-1	mAP	R-1	mAP
				66.3	57.1	81.5	77.8
✓				68.1	59.8	83.7	80.9
✓	✓			69.2	60.7	84.9	82.0
✓	✓	✓		71.8	64.3	86.7	83.7
✓	✓	✓	✓	72.8	65.2	88.2	84.0

TABLE IV
THE EFFECTIVENESS OF OUR PROPOSED MODULES ON THE OCCLUDED-DUKE (O-DUKE) AND OCCLUDED-REID (O-REID) DATASETS. THE "CONTRASTATTN" REPRESENTS OUR CONTRASTIVE PEDESTRIAN ATTENTION MODULE, IN WHICH WE EMPLOY AN INTER-IMAGE CONTRASTIVE LOSS (INTER.) AND AN INTRA-IMAGE CONTRASTIVE LOSS (INTRA.) TO TRAIN THIS MODULE. THE "CORREL" REPRESENTS OUR CORRELATION LEARNING MODULE, IN WHICH WE EMPLOY THE 4D CONVOLUTIONAL BLOCKS (4D CONVS) TO CAPTURE CORRESPONDENCES BETWEEN COMPARING IMAGES. BESIDES, IN THIS MODULE, A POSITION EMBEDDING (POSEMB.) IS GIVEN TO ENABLE THE 4D CONVS TO LEVERAGE THE SPATIAL INFORMATION OF COMPARING PEDESTRIAN CLUES TO CAPTURE CORRESPONDENCES MORE EFFECTIVELY.

D. Ablation Studies

In this subsection, we conduct a series of ablation experiments on the Occluded-Duke (O-Duke) and Occluded-REID (O-REID) datasets to verify the effectiveness of each module in our proposed CpaCol. Noting that the "baseline" model mentioned in the rest of this manuscript indicates the TransReID model [30], which utilizes ViT [29] as the backbone model and is trained with identity classification loss and triplet loss.

Effectiveness of Our Proposed Modules. To evaluate the effectiveness of our contrastive pedestrian attention (ContrastAttn) module and correlation learning (CorreL) module, we gradually add them to the baseline model and compare the performance improvements on the O-Duke and O-REID datasets. The experimental results are exhibited in Table IV. Besides, in our ContrastAttn module, we have developed two forms of contrastive loss: inter-image contrastive loss (Inter.) and intra-image contrastive loss (Intra.). These two losses prompt our model to learn the semantic divergence between

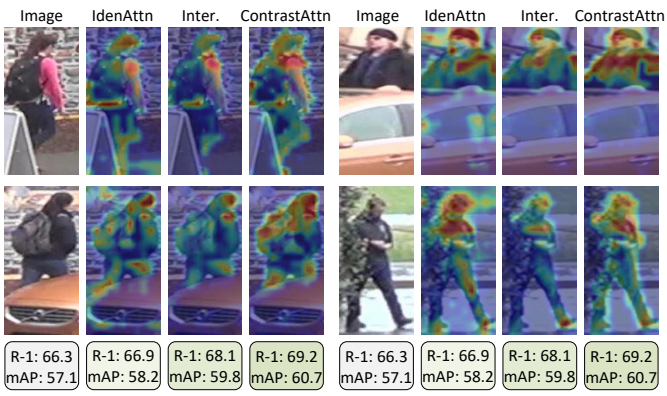

Fig. 7. The experimental comparisons between the existing identity-guided attention mechanism (IdenAttn) and our designed contrastive pedestrian attention module (ContrastAttn) on the O-Duke dataset. Here, to respectively explore the effectiveness of our designed inter-image and intra-image contrastive losses, the performance of version only employing inter-image contrastive loss (Inter.) is also given. For comparison, the results of the baseline model are listed in the "Baseline" column. From the visualization, we can find that only supervising the final predictions with identity-level losses could mislead the IdenAttn to treat some pedestrian-irrelevant noises as discriminative clues for re-identification. In comparison, the version employing inter-image contrastive loss could filter out these noises and focus on pedestrian contents, thereby achieving 1.6% mAP improvement. Besides, the additional intra-image contrastive loss could help our ContrastAttn to capture pedestrian foregrounds more comprehensively, thus further bringing 0.9% mAP improvement.

pedestrian foregrounds and pedestrian-irrelevant noises, as well as the semantic consistency across different human body parts, respectively. Here, we separately integrate these two losses into the baseline model to investigate their effectiveness. As depicted in Table IV, the inter-image contrastive loss contributes to an average R-1 improvement of 2.0% over the baseline model, while the intra-image contrastive loss offers an additional 1.2% R-1 improvement on average. These enhancements indicate the effectiveness of our contrastive learning strategy. Additionally, we further incorporate the CorreL module and assess the improvement it brings. Here, we also evaluate the effectiveness of our proposed position embedding (PosEmb.) by comparing the performance of versions with and without the position embedding. From the results in Table IV, we can find that using original 4D convolution blocks (4D Convs) to capture correspondences between comparing images could averagely bring 2.2% of R-1 improvement. Besides, using our proposed position embedding, the CorreL module could capture the correspondences more effectively, thus further achieving an average of 1.3% of R-1 improvement. These improvements suggest that our employed 4D convolutional blocks and position embedding are both effective in capturing corresponding clues between comparing images, thereby improving re-identification.

Effectiveness of Our Contrastive Learning Strategy. In this part, we aim to investigate the efficacy of our contrastive learning strategy in enhancing the attention mechanism. For this purpose, we respectively incorporate our ContrastAttn module and the identity-guided attention module (removing the inter-image and intra-image contrastive losses, denoted as IdenAttn) into the baseline model, comparing their performances on the O-Duke dataset, as shown in Fig. 7. Besides, to further explore the efficacy of our proposed inter-image contrastive loss and intra-image contrastive loss separately, we also give the performance of the ContrastAttn module solely using the inter-image contrastive loss (Inter.).

As demonstrated in the "IdenAttn" column, we can find that despite the identity-guided attention module contributing an mAP improvement of 1.1% to the baseline model (the first column), there still exists a 2.5% mAP gap between it and the version incorporating our ContrastAttn module. This performance gap may stem from the fact that the IdenAttn module only supervises final predictions via identity-level loss, neglecting the causality between its generated attention maps and the subsequent predictions. This could mislead it to capture some discriminative yet pedestrian-irrelevant noises, thus impacting re-identification. For example, in the right case at the first row of Fig. 7, the IdenAttn module incorrectly considers the occluded vehicle as a significant clue, thus focusing on it. In contrast, under the guidance of our inter-image contrastive loss, the ReID model could learn the divergence between the pedestrian foregrounds and noises, thereby enabling it to capture pedestrian foregrounds more accurately. Specifically, comparing the visualization results in the "IdenAttn" and "Inter." columns, we could find that our inter-image contrastive loss effectively prevents the ReID model from capturing pedestrian-irrelevant noises, leading to a 1.6% mAP improvement. Furthermore, comparing the results in the "Inter." and "ContrastAttn" columns, we can find that incorporating an additional intra-image contrastive loss could prompt our ContrastAttn module to capture the pedestrian foregrounds more comprehensively, thereby realizing a further 0.9% mAP improvement.

Dataset	Training Strategy	R-1	R-5	R-10	mAP
O-Duke	End-to-end	69.3	83.9	88.0	62.8
	Two-stage	**72.8**	**85.1**	**89.5**	**65.2**
O-REID	End-to-end	85.0	94.0	96.1	81.3
	Two-stage	**88.2**	**95.5**	**97.4**	**84.0**

TABLE V
COMPARISON RESULTS OF THE VERSIONS EMPLOYING OUR TWO-STAGE TRAINING STRATEGY OR TRAINING THE PEDESTRIAN FEATURE EXTRACTION MODEL AND THE CORREL MODULE FROM SCRATCH TOGETHER IN AN END-TO-END MANNER.

Effectiveness of Two-Stage Training Strategy. During the training phase, we adopt a two-stage strategy to train our CapCol model. The rationale behind this strategy is that the 4D convolution [23] inside our correlation learning (CorreL) module needs to employ features extracted by a well-trained extractor to capture correspondences between comparing images. Hence, we consider first thoroughly training the pedestrian feature extraction model (ViT backbone inserted with our ContrastAttn module) and then using its extracted features to train our CorreL module. Here, we aim to explore the effectiveness and necessity of our two-stage training strategy. To do so, we directly train the pedestrian feature extraction model and the CorreL module from scratch together in an end-to-end manner (End-to-end), and then compare it with the version using our two-stage training strategy. The comparison results are exhibited in Table V. From these results, we can

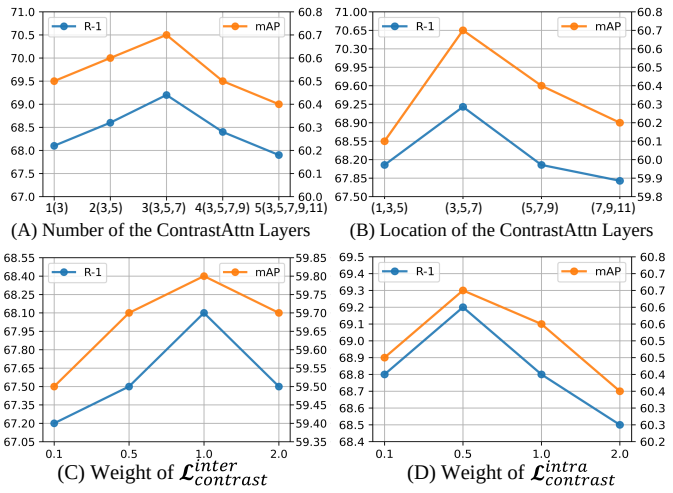

Fig. 8. Ablation experiments about (A) the number of employed contrastive pedestrian attention (ContrastAttn) layers, (B) the locations to insert ContrastAttn layers, (C) the loss weight of inter-image contrastive loss, and (D) the loss weight of intra-image contrastive loss. All these ablation experiments are conducted on O-Duke. In the X-axis annotations of (A) and (B), the numbers in "(,)" indicates which layers of ViT we insert the ContrastAttn layers into. The left and right parts of the Y-axis scales indicate rank-1 accuracy (R-1) and mAP, respectively.

find that our two-stage training strategy is indeed necessary, significantly outperforming the version training these two modules end-to-end by $2.4/2.7\%$ mAP on O-Duke and O-REID, respectively.

Ablation about Deployment and Hyper-parameters. We give some ablation studies about the deployment and some crucial hyper-parameters of our CpaCol model in Fig. 8. Firstly, we would like to explore the most suitable locations of ViT to insert the contrastive pedestrian attention (ContrastAttn) layers. To do so, we perform some quantitative experiments on the O-Duke dataset. In Fig. 8, we exhibit the R-1 and mAP performance of versions inserting different numbers of ContrastAttn layers in (A) and versions inserting them into different locations in (B). As shown in Fig. 8 (A), the performance gradually improves as the number of inserted ContrastAttn layers increases from 1 to 3, while decreasing if more than 3 ContrastAttn layers are utilized. In addition, to further explore the influence of the inserted location, we fix the number of employed ContrastAttn layers to 3 and insert them into the low (1,3,5), middle (3,5,7), (5,7,9), and high (7,9,11) layers of ViT, respectively. The results depicted in Fig. 8 (B) indicate that inserting the ContrastAttn layers into the (3,5,7) layers yields superior performance. The likely explanation for this finding could be that the ReID fails to extract sufficient semantic information in the low layers, such as the first layer, to effectively differentiate between pedestrian foregrounds and noises. On the other hand, when our ContrastAttn module is applied to the middle layers, *i.e.*, (3,5,7), it is able to access a more substantial amount of semantic information and could remove noises early enough, which subsequently prompts the following layers to extract pedestrian foreground features. Consequently, in our experiments, we insert three ContrastAttn layers into the 3rd, 5th, and 7th layers of the ViT model.

Model	R-1	mAP	GFLOPs	Params(M)
Baseline	66.3	57.1	22.52	92.74
+ ContrastAttn	69.2	60.7	$23.60^{\uparrow 1.08}$	$98.82^{\uparrow 6.08}$
CpaCol	72.8	65.2	$24.89^{\uparrow 1.29}$	$104.13^{\uparrow 5.31}$

TABLE VI
ABLATION STUDIES ABOUT THE COMPUTATIONAL COMPUTATION AND CAPACITY OF OUR DESIGNED MODULES. THE EXPERIMENTS ARE CONDUCTED ON THE O-DUKE DATASET.

Secondly, we also explore the optimal loss weight for our employed inter-image contrastive loss $\mathcal{L}_{contrast}^{inter}$ and intra-image contrastive loss $\mathcal{L}_{contrast}^{intra}$, which are respectively responsible for guiding our model to learn the semantic distinction between pedestrian foregrounds and noises as well as the consistency between different pedestrian parts. In this step, we first especially explore the best weight for $\mathcal{L}_{contrast}^{inter}$ by removing $\mathcal{L}_{contrast}^{intra}$. As shown in Fig. 8 (C), we can find that our model achieves the best performance when the weight is set to 1.0. This finding might be attributed to the fact that a small loss weight, *e.g.*, 0.1 and 0.5, may be insufficient to guide our model in distinguishing pedestrian foregrounds from background noises. On the other hand, an excessively large loss weight, such as 2.0, could over-force our model to capture the divergence between foregrounds and noises, while potentially hindering it from distinguishing the foregrounds of different pedestrian instances. In light of this finding, we fix the weight of $\mathcal{L}_{contrast}^{inter}$ to 1.0 and proceed to explore the optimal weight for $\mathcal{L}_{contrast}^{intra}$. As depicted in Fig. 8 (D), our model performs optimally when the weight is set to 0.5. Therefore, in this work, we set the weights of $\mathcal{L}_{contrast}^{inter}$ and $\mathcal{L}_{contrast}^{intra}$ to 1.0 and 0.5, respectively.

Analysis about Model Capacity. In this study, we develop a ContrastAttn module aiming to capture pedestrian foregrounds, as well as a CorreL module to identify correspondences between comparing images. Here, we aim to explore the computational consumption associated with our proposed modules. To accomplish this, we gradually incorporate these two modules into the baseline model, subsequently exploring both the additional computational consumption and the performance improvement they bring. The comparative results are presented in Table VI. From these results, we can observe that our proposed modules augment the computational consumption by 1.08 and 1.29 GFLOPs, respectively. Since these increases are an order of magnitude lower than the total computation of the baseline model (22.52 GFLOPs), it is reasonable to render them acceptable. Furthermore, it's important to note that our designed modules both deliver substantial performance improvements, specifically, 3.6% and 4.5% increases in mAP. This highlights the necessity of incorporating them into our model.

E. Retrieval Cases

To intuitively illustrate how our designed correlation learning (CorreL) module mitigates the impact caused by information asymmetry, we provide several retrieval cases in Fig. 9.

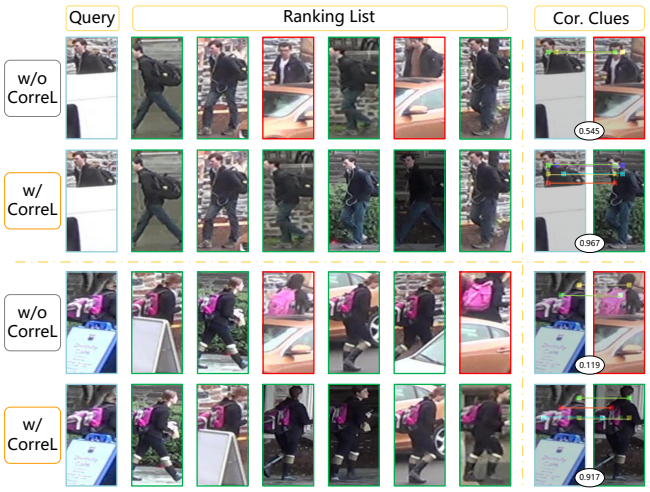

Fig. 9. Retrieval cases of the pedestrian feature extraction model ("w/o CorreL") and the version incorporating correlation learning module ("w/ CorreL"). Besides, the right part exhibits the corresponding clues (Cor. Clues) captured by the CorreL module and the integrated image-level similarity between the confusing image pairs. We can find that our CorreL module could capture rich corresponding clues between paired occluded and holistic images, thereby predicting a high image-level similarity for them. This could help our model focus more on the correspondences between compared images, avoiding biased predictions. Conversely, for non-corresponding images, our CorreL module identifies only a few correspondences, predicting a relatively low similarity and subsequently reducing their ranking orders.

These cases compare the retrieval results of the pedestrian feature extraction model (a ViT backbone integrated with our ContrastAttn module, donated as "w/o CorreL") and the version incorporating our correlation learning module ("w/ CorreL"). As shown in the "w/o CorreL" rows of Fig. 9, due to the semantic gap caused by information asymmetry between occluded and holistic images, the vanilla pedestrian feature extraction model erroneously underestimates the similarity between the occluded query image and its corresponding holistic images. This misjudgment leads it to match the occluded query image with some non-corresponding occluded images. In contrast, as depicted in the "w/ CorreL" rows, the model employing our CorreL module could effectively alleviate the interference caused by information asymmetry and match the corresponding occluded and holistic images.

To better understand how our module facilitates the matching of paired images, in the right part of Fig. 9, we provide visualizations of the corresponding clues captured by our CorreL module and the integrated image-level similarities, *i.e.*, $s^{i,j}$ in Eq. 10, which would be aggregated with the global-level similarity together for prediction. These visualizations show that our CorreL module could capture rich corresponding clues between paired occluded and holistic images, thereby predicting a high image-level similarity for them. This assists our model in focusing more on the correspondences between compared images, avoiding biased predictions. Conversely, for non-corresponding images, our CorreL module identifies only a few correspondences, predicting a relatively low similarity and reducing their ranking orders.

V. Conclusion

In this paper, we propose a contrastive pedestrian attentive and correlation learning (CpaCol) model for occluded person ReID. The CpaCol consists of two major components, *i.e.*, a contrastive pedestrian attention (ContrastAttn) module and a correlation learning (CorreL) module. Specifically, in the ContrastAttn module, a contrastive learning strategy is adopted to guide the attention module to learn the divergence between pedestrian foregrounds and noises, thereby enabling it to capture foregrounds effectively. Besides, in the CorreL module, we employ 4D convolution to help the ReID model to capture and focus more on the fine-grained correspondences between comparing images. This could prevent our model from overemphasizing the inherent global-level semantic divergence between occluded and holistic images, thereby mitigating the risk of biased predictions. Extensive experimental results demonstrate the effectiveness of our proposed modules.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[2] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding DINO: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[3] 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. Comp. Vis.*, 2017, pp. 3754–3762.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 770–778.

[5] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 365–378, 2018.

[6] ——, "Robust structural sparse tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 473–486, 2018.

[7] Z. Li, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 1239–1248.

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

[9] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 402–419.

[10] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 420–429.

[11] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 4099–4108.

[12] Q. Zhou, H. Fan, S. Zheng, H. Su, X. Li, S. Wu, and H. Ling, "Graph correspondence transfer for person re-identification," in *Proc. Conf. AAAI*, vol. 32, no. 1, 2018, pp. 7599–7606.

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

[14] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2019, pp. 542–551.

[15] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person reid," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 11 744–11 752.

[16] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 6449–6458.

[17] H. Huang, X. Chen, and K. Huang, "Human parsing based alignment with multi-task learning for occluded person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo*. IEEE, 2020, pp. 1–6.

[18] 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 Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 2898–2907.

[19] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q. Chen, and R. Ji, "Occlude them all: Occlusion-aware attention network for occluded person Re-ID," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 11 833–11 842.

[20] B. Xu, L. He, J. Liang, and Z. Sun, "Learning feature recovery transformer for occluded person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 4651–4662, 2022.

[21] G. Yan, Z. Wang, S. Geng, Y. Yu, and Y. Guo, "Part-based representation enhancement for occluded person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[22] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 6941–6952.

[23] S. Lee, H. Seong, S. Lee, and E. Kim, "Correlation verification for image retrieval," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 5374–5384.

[24] B. Jiao, L. Liu, L. Gao, G. Lin, R. Wu, S. Zhang, P. Wang, and Y. Zhang, "Generalizable person re-identification via viewpoint alignment and fusion," *arXiv preprint arXiv:2212.02398*, 2022.

[25] B. Jiao, X. Tan, J. Zhou, L. Yang, Y. Wang, and P. Wang, "Instance and pair-aware dynamic networks for re-identification," *arXiv preprint arXiv:2103.05395*, 2021.

[26] B. Jiao, L. Liu, L. Gao, G. Lin, L. Yang, S. Zhang, P. Wang, and Y. Zhang, "Dynamically transformed instance normalization network for generalizable person re-identification," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2022, pp. 285–301.

[27] B. Jiao, L. Gao, and P. Wang, "Temporal-consistent visual clue attentive network for video-based person re-identification," in *ACM Int. Conf. Multimedia Retrieval*, 2022, pp. 72–80.

[28] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2019, pp. 3702–3712.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.

[30] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 15 013–15 022.

[31] 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 *Proc. IEEE Int. Conf. Comp. Vis.*, 2019, pp. 3642–3651.

[32] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval," in *ACM Int. Conf. Multimedia*, 2017, pp. 420–428.

[33] 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 Conf. Comp. Vis. Patt. Recogn. Workshops*, 2019.

[34] 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. Comp. Vis.*, 2018, pp. 480–496.

[35] Y. Huang, Y. Huang, H. Hu, D. Chen, and T. Su, "Deeply associative two-stage representations learning based on labels interval extension loss and group loss for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4526–4539, 2019.

[36] 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, 2021.

[37] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019, pp. 9317–9326.

[38] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019, pp. 1389–1398.

[39] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 4692–4702.

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 2921–2929.

[41] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proc. Conf. AAAI*, vol. 36, no. 3, 2022, pp. 2540–2549.

[42] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 4754–4763.

[43] H. Tan, X. Liu, B. Yin, and X. Li, "MHSA-Net: Multihead self-attention network for occluded person re-identification." *IEEE Trans. Neural Netw. & Learn. Syst.*, 2022.

[44] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 7291–7299.

[45] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 33, pp. 18 661–18 673, 2020.

[46] T. Barletti, N. Biondi, F. Pernici, M. Bruni, and A. Del Bimbo, "Contrastive supervised distillation for continual representation learning," in *Int. Conf. Image Anal. Process.*, 2022, pp. 597–609.

[47] J. Li, P. Zhou, C. Xiong, and S. Hoi, "Prototypical contrastive learning of unsupervised representations," in *Proc. Int. Conf. Learn. Representations*, 2021.

[48] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 9729–9738.

[49] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 1179–1188.

[50] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint generative and contrastive learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 2004–2013.

[51] S. Choudhury, I. Laina, C. Rupprecht, and A. Vedaldi, "Unsupervised part discovery from contrastive reconstruction," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 34, pp. 28 104–28 118, 2021.

[52] S. Li, K. Han, T. W. Costain, H. Howard-Jenkins, and V. Prisacariu, "Correspondence networks with adaptive neighbourhood consensus," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 10 196–10 205.

[53] J. Min and M. Cho, "Convolutional hough matching networks," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 2940–2950.

[54] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 31, pp. 1651–1662, 2018.

[55] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[56] G. Yang and D. Ramanan, "Volumetric correspondence networks for optical flow," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 32, 2019.

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

[58] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2011, pp. 649–656.

[59] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2014, pp. 688–703.

[60] 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 Int. Conf. Comp. Vis.*, 2019, pp. 8450–8459.

[61] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. Conf. AAAI*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[62] 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, 2020.

Liying Gao received the B.E. degree from Northwestern Polytechnical University in 2018. Now she is pursuing the Ph.D. degree at Northwestern Polytechnical University. Her current research interests include deep learning and image-text matching.

Bingliang Jiao received the B.E. degree from Northwestern Polytechnical University in 2018. Now he is pursuing the Ph.D. degree at Northwestern Polytechnical University. His current research interests include deep learning, Image- retrieval and ReID.

Yuzhou Long received the B.E. degree from Northwestern Polytechnical University in 2021. Now he is pursuing the master's degree at Northwestern Polytechnical University. His current research interests include deep learning and image-text retrieval.

Kai Niu received the B.E. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2015, and the Ph.D. degree from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2020. He is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University. His main research interests include multi-modal computing, computer vision, and machine learning.

He Huang received his doctor's degree in the college of Astronautics, Northwestern Polytechnical University (NPU) in 2010. He was a visiting scholar in Technical University of Munich Germany from December 2016 to December 2017. He is an associate professor in the school of Astronautics, NPU. His research interests are including: space situational awareness, microsatellite system design, satellite dynamic and control, satellite formation and swarm flight control.

Peng Wang received the B.E. degree in electrical engineering and automation and received the Ph.D. degree in control science and engineering from Beihang University, in 2004 and 2011, respectively. He is currently a professor in the School of Computer Science at Northwestern Polytechnical University. His current research interests include computer vision, machine learning and artificial intelligence.

Yanning Zhang received the Ph.D. degree from the School of Marine Engineering, Northwestern Polytechnical University in 1996. She is currently a Professor at the School of Computer Science, Northwestern Polytechnical University. Her current research interests include computer vision and pattern recognition, image and video processing, and intelligent information processing.