Completed Part Transformer for Person Re-identification

Zhong Zhang, *Senior Member, IEEE,* Di He, Shuang Liu, *Senior Member, IEEE*, Baihua Xiao, Tariq S. Durrani, *Life Fellow, IEEE*

Abstract—**Recently, part information of pedes
been demonstrated to be effective for person
(ReID), but the part interaction is ignored wl
former to learn long-range dependencies. In
propose a novel transformer network named
Transformer (CPT) for person ReID, where we
transformer layer to learn the completed part
part transformer layer includes the intra-par
part-global layer, where they consider long-ran
from the aspects of the intra-part interactio
global interaction, simultaneously. Furthermo
overcome the limitation of fixed number of the
the transformer layer, we propose the Adaptive
(ART) module to focus on learning the inte
the informative patch tokens in the pedestria
improves the discrimination of the pedestrian
Extensive experimental results on four person
i.e., MSMT17, Market1501, DukeMTMC-reII
demonstrate that the proposed method achieves a new state-of-the-art performance, e.g., it achieves 68.0% mAP and 84.6% Rank-1 accuracy on MSMT17.**

Index Terms—**Person ReID, Transformer, Adaptive Refined Tokens.**

I. Introduction

PERSON re-identification (ReID) [1]–[4] aims to associate the target pedestrian across multiple non-overlapping cameras, which has become an important component in the intelligent video surveillance system. However, it is a challenging task because pedestrian images are captured from unconstrained environments, where the pedestrian appearances are easily influenced by many factors such as occlusions, illuminations, viewpoints, poses, etc [5]–[7].

In order to overcome the above-mentioned challenges, a large number of studies focus on learning robust and discriminative features for person ReID [8]–[11]. With the rapid development of deep learning, deep feature learning methods [2], [11]–[14] have significantly improved the performance of the person ReID task. However, most of them usually adopt the Convolutional Neural Networks (CNNs) as a backbone to extract deep features, where the long-range dependencies are neglected. This is because the CNNs-based methods are mainly composed of a series of convolution filters having limited receptive field sizes. The long-range dependencies are also known as the global context dependencies, which are mainly represented by the interactive learning from the whole pedestrian image.

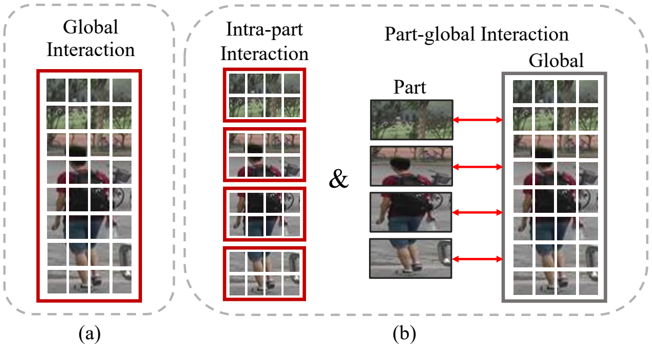


Fig. 1. (a) The patch tokens are obtained from the whole pedestrian image. (b) The proposed method generates patch tokens from the stripe parts of pedestrian image, where we build both intra-part interaction and part-global interaction so as to learn the completed part interaction. The red box and the red bidirectional arrow indicates to learn the interaction using self-attention and the cross-attention, respectively.

Recently, Transformer [15] is proposed to learn long-range dependencies using self-attention mechanism, and it is successfully applied in Natural Language Processing (NLP). Afterwards, Vision Transformer (ViT) [16] as a pure transformer model achieves huge success in image classification, and then it is rapidly expanded to many vision tasks, such as semantic segmentation [17], visual object tracking [18] and GANs [19]. As for person ReID, some transformer methods [9], [20], [21] divide pedestrian image or feature maps into a series of non-overlapping patch tokens to learn global interaction as shown in Fig. 1 (a). Here, each patch token is changed into a feature vector, and the interaction between the patch tokens simultaneously represents the interaction between feature vectors, hence learning the relationships of the patch tokens in different positions using the self-attention mechanism. However, there are two limitations of the transformer methods for person ReID. Firstly, the information of rigid stripe parts has been

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62171321, in part by the Natural Science Foundation of Tianjin under Grant No. 22JCQNJC00010 and No. 20JCZDJC00180, in part by the Scientific Research Project of Tianjin Educational Committee under Grant No. 2022KJ011, and the Key Project of Graduate Scientific Research Innovation Project of Tianjin Normal University under Grant No. 2022KYCX032Z. (Corresponding author: Shuang Liu.)

Zhong Zhang, Di He and Shuang Liu are with Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China (e-mail: {zhong.zhang8848, clarkhedi, shuangliu.tjnu}@gmail.com).

Baihua Xiao is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: baihua.xiao@ia.ac.cn).

Tariq S. Durrani is with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow Scotland, UK (e-mail: t.durrani@strath.ac.uk).

proven to be effective for person ReID, but it is ignored when learning the interaction between patch tokens. Secondly, the number of patch tokens in the transformer layer keeps fixed, which leads to insufficient learning of salient regions for pedestrian images. Hence, the transformer model needs to be designed for person ReID so as to overcome these shortcomings.

In this paper, we propose a novel transformer network named Completed Part Transformer (CPT) for person ReID, where we design the part transformer layer to learn the completed part interaction. The part transformer layer could learn long-range dependencies from the aspects of the intra-part interaction and the part-global interaction, and it consists of the intra-part layer and the part-global layer. Specifically, as for the intra-part layer, we first split each pedestrian image into several stripes, in which each stripe is divided and mapped into a series of non-overlapping patch tokens as shown in the left of Fig. 1 (b). Then, it builds long-range dependencies by learning the interaction between all patch tokens in each stripe. Unlike the traditional global interaction, we focus on the interaction between the patch tokens within each part so as to integrate the part information when learning the interaction.

In order to learn the completed part interaction, we further learn the interaction between the stripe part and the patch tokens of the whole pedestrian image so as to make the stripe part incorporate more information. Correspondingly, we propose the part-global layer to obtain the part-global interaction. Specifically, we learn the interaction between the class tokens of the stripes and the output of the previous part-global layer to obtain the fused class tokens. Hence, the fused class tokens could represent the part information more flexibly due to interacting the learnable information. Afterwards, we build the part-global interaction between the fused class tokens and the patch tokens from the whole pedestrian image by applying the cross-attention mechanism as shown in the right of Fig. 1 (b), where we take the fused class tokens as queries and the patch tokens from the whole pedestrian image as keys and values. Finally, we aggregate the output of the last part-global layer through the max pooling to obtain the global token for subsequent optimization.

Furthermore, the fine-grained information of human body parts is a vital clue to distinguish the pedestrian, and meanwhile some patch tokens in each stripe contain irrelevant information of pedestrian images. But, the fixed number of patch tokens in the transformer layer is difficult to mine fine-grained information and restrain irrelevant information for person ReID. Hence, we propose the Adaptive Refined Tokens (ART) module to retain the patch tokens with more information in each stripe during the learning process of the intra-part interaction. Since the patch tokens learned by the part transformer layer contain the completed part information, we generate the part masks based on the affinity matrices of the part transformer layer. Then, we insert the part masks into the intra-part layer to select the patch tokens with more information in each stripe and build their dependencies in order to learn discriminative features.

The main contributions of the proposed method are summarized as follows:

- We propose CPT to learn the completed part interaction for person ReID, where we design the part transformer layer to consider long-range dependencies from the aspects of the intra-part interaction and the part-global interaction.
- We propose the ART module to improve the discrimination of the pedestrian features by retaining informative patch tokens, where we utilize the part masks to establish the dependencies of the patch tokens with more information in each stripe of pedestrian image.
- Extensive experimental results demonstrate that the proposed method achieves a new state-of-the-art performance on four person ReID datasets.

II. Related Work

In this section, we first introduce the person ReID, and then introduce the visual transformer.

A. Person ReID

The existing person ReID methods are mainly divided into three categories, i.e., hand-crafted descriptors [22], [23], metric learning methods [24]–[28] and deep learning methods [2], [8], [11], [13], [29]. In recent years, deep learning methods have become mainstream approaches in the person ReID task due to the promising performance. These approaches focus on learning global features [11], [30]–[34] and local features [1], [2], [12], [13], [35], [36].

As for the global features, Wang *et al.* [30] propose a joint learning framework consisting of single-image global representation and cross-image global representation for person ReID. Chen *et al.* [31] propose to jointly optimize classification tasks and ranking tasks simultaneously so as to learn discriminative global features for person ReID. Li *et al.* [11] propose Pose-Guided Representation (PGR) learning for person ReID, where they consider the human part cues to supervise the training process of global features.

The local features of pedestrian images could offer body structure information which is beneficial for person ReID. For example, Sun *et al.* [2] propose to learn stripe-based local features by dividing pedestrian images into fixed horizontal stripes. Wang *et al.* [12] split pedestrian images into overlapping stripes with different granularities so as to learn multi-scale local features. Zhang *et al.* [13] propose Heterogeneous Local Graph Attention Networks (HLGAT) for person ReID, in which they learn the intra-local relation and the inter-local relation by modeling the completed local graph. Ding *et al.* [36] propose the Multi-task Part-aware Network (MPN) which is designed to extract semantically aligned part-level features from pedestrian images for person ReID.

Furthermore, some person ReID methods [37]–[40] utilize the attention mechanism to enhance the representations of pedestrian images. For example, Wang *et al.* [38] propose the fully attentional block which creates both channel-wise and spatial-wise attention information to deal with the misalignment problem and localize discriminative local features so as to mine the useful features. Chen *et al.* [39] propose the self-critical attention learning method to improve the effectiveness

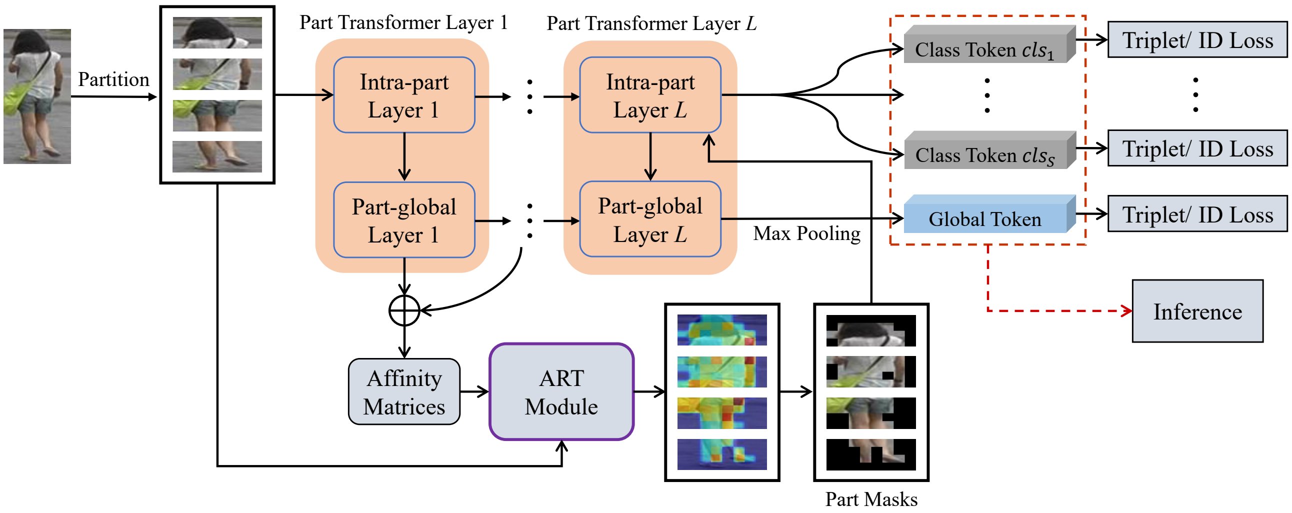


Fig. 2. The overall framework of the proposed method for person ReID. We first split each pedestrian image into several stripes, and then feed them into L part transformer layers where each part transformer layer consists of the intra-part layer and the part-global layer. Furthermore, we propose the ART module to generate the part masks based on the affinity matrices of the part-global layers, and we insert the part masks into the intra-part layer to retain informative patch tokens in each stripe. In the test stage, we employ the global token from the part-global layer and the class tokens from the intra-part layer for inference.

of the attention model by considering the attention confidence level. Ren *et al.* [40] build an end-to-end network S^2-Net which designs the semantic attention to learn the human semantic partition and the saliency attention to capture the salient non-human parts.

B. Visual Transformer

In recent times, the transformer model has been widely used in computer vision fields [16], [41], [42]. Due to its powerful modeling capability, the transformer model is applied in the field of person ReID. He *et al.* [9] propose a pure transformer model to learn discriminative features by using the side information embedding and a jigsaw patch module, and achieve promising performance on person ReID.

Furthermore, some methods [20], [21], [43]–[47] combine CNNs and the transformer model for person ReID. For example, Liao *et al.* [21] propose TransMatcher to consider image-to-image attention via the encoder-decoder transformer architecture, where the first transformer layer is replaced by CNNs. Li *et al.* [43] propose Part Aware Transformer (PAT) to learn robust human part discovery by combining both the CNNs backbone and a transformer encoder-decoder architecture. Zhang *et al.* [20] present Hierarchical Aggregation Transformers (HAT) to learn multi-scale features by embedding the transformer model into each layer of CNNs. Wang *et al.* [46] exploit CNNs to learn pose information of pedestrian, and then disentangle semantic components using the transformer model. Wang *et al.* [47] propose Neighbor Transformer Network (NFormer) to explicitly model the interaction across all images, where promising results are obtained by using ABDNet [48] as the backbone.

Different from the above-mentioned methods, we propose CPT to learn the completed part interaction via the well-designed part transformer layer for person ReID. Furthermore, we propose the ART module to improve the discrimination of the pedestrian features by retaining the patch tokens with more information.

III. Approach

In this section, we initially review the mechanism of ViT, and then describe the major parts of the proposed method, i.e., the part transformer layer and the ART module in detail.

A. ViT Revisit

Given a pedestrian image $X \in \mathbb{R}^{H\times W\times C}$, some methods [9], [16] divide the image into a sequence of patches $\{x_i \in \mathbb{R}^{1\times K^2\cdot C} | i = 1, \ldots, N\}$, where C, H and W denote the number of channels, height and width of pedestrian image, respectively. Here, $K \times K$ is the size of image patch, and $N = HW/K^2$ is the number of patches. Then, these patches are mapped as the patch tokens using a trainable linear projection $E(\cdot)$ implemented by a fully-connected layer with the neuron number D. Finally, these patch tokens are fed into L transformer layers, and the input of the first transformer layer is represented as:

$$Z^0 = \mathrm{cat1}(cls, E(x_1), \cdots, E(x_N)) + \mathcal{P} + \lambda J, \tag{1}$$

where $\mathrm{cat1}$ indicates the concatenation operation along the column, $cls \in \mathbb{R}^{1\times D}$ denotes the learnable class token which retains the information of the pedestrian image by interacting with all patch tokens, $\mathcal{P} \in \mathbb{R}^{(N+1)\times D}$ is the learnable position embedding, $J \in \mathbb{R}^{(N+1)\times D}$ is the camera embedding, and λ is a parameter to balance the weight of the camera embedding. After multiple transformer layers, the output of the l-th transformer layer is formulated as:

$$Z^l = \hat{Z}^{l-1} + MLP(LN(\hat{Z}^{l-1})), \tag{2}$$

$$\hat{Z}^{l-1} = Z^{l-1} + MSA(LN(Z^{l-1})), \tag{3}$$

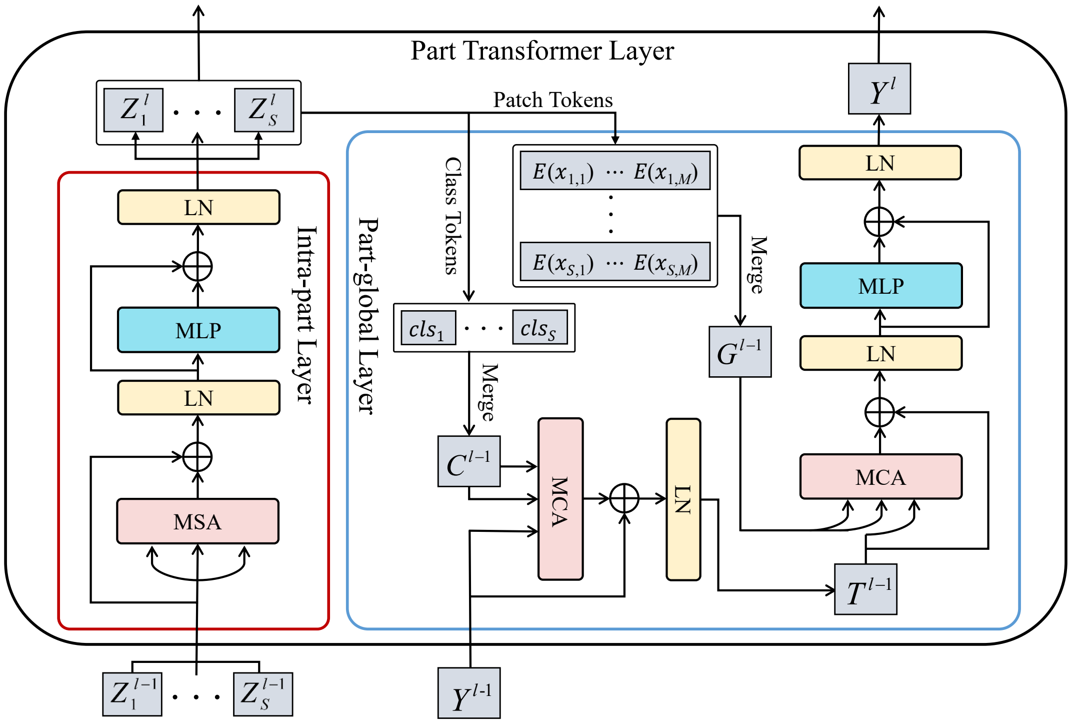


Fig. 3. The structure of the part transformer layer. When $l = 1$, $Y^{l-1} \in \mathbb{R}^{S \times D}$ indicates the learnable parameters.

where LN, MLP and MSA denote the layer normalization, the multi-layer perceptron and the multi-head self-attention, respectively. Specifically, the multi-head self-attention is defined as:

$$MSA(Z^{l-1}) = \text{cat2}(\alpha_1^l V_1^l, \cdots, \alpha_h^l V_h^l, \cdots, \alpha_H^l V_H^l)U^l, \quad (4)$$

$$V_h^l = Z^{l-1} W_v^{l,h}, \quad (5)$$

where cat2 indicates the concatenation operation along the row, H indicates the number of heads, α_h^l is the affinity matrix of the h-th head in the l-th transformer layer, and $W_v^{l,h} \in \mathbb{R}^{D \times d}$ and $U^l \in \mathbb{R}^{D \times D}$ are the linear projections. Here, the affinity matrix is defined as:

$$\alpha_h^l = Softmax(\frac{Q_h^l (K_h^l)^T}{\sqrt{d}}), \quad (6)$$

$$Q_h^l = Z^{l-1} W_q^{l,h}, K_h^l = Z^{l-1} W_k^{l,h}, \quad (7)$$

where $d = D/H$, $\sqrt{d}$ is used to normalize for numerical stability, and $W_q^{l,h} \in \mathbb{R}^{D \times d}$ and $W_k^{l,h} \in \mathbb{R}^{D \times d}$ are the linear projections.

B. Part Transformer Layer

The part information has exhibited effective results for person ReID [1], [2], [8], [13], however the existing transformer methods cannot explicitly learn the part information of pedestrian [9], [20], [47]. To overcome this drawback, we propose CPT to learn the completed part interaction via the well-designed part transformer layer for person ReID. The framework of our method is shown in Fig. 2. We first split each pedestrian image into several stripes, and then learn long-range dependencies from the aspects of the intra-part interaction and the part-global interaction via L part transformer layers. Here, the part transformer layer consists of the intra-part layer and the part-global layer.

Intra-part Layer. Since the stripes indicate the spatial distribution of human parts, we propose the intra-part layer to learn the intra-part interaction as shown in the red box of Fig. 3. The intra-part interaction represents the interaction between the patch tokens in each stripe part, where each patch token is a feature vector and the patch tokens in one stripe part learn from each other by using self-attention mechanism so as to mine the long-range dependencies within stripe parts of pedestrian images.

Before feeding the intra-part layer, we first divide each pedestrian image $X \in \mathbb{R}^{H \times W \times C}$ into S stripes, where each stripe is also partitioned into a series of patches $\{x_{s,i} \in \mathbb{R}^{1 \times K^2 \cdot C} | s = 1, 2, \ldots, S; i = 1, \ldots, M\}$ using a sliding window with non-overlapping pixels. Here, $M = N/S$ is the number of patches in a stripe. Subsequently, similar to Eq. 1, the input of the first intra-part layer for the s-th stripe is formulated as:

$$Z_s^0 = \text{cat1}(cls_s, E(x_{s,1}), \cdots, E(x_{s,M})) + \mathcal{P}_s + \lambda J_s, \quad (8)$$

where $cls_s \in \mathbb{R}^{1 \times D}$ denotes the learnable class token of the s-th stripe which maintains the information of the stripe, $\mathcal{P}_s \in \mathbb{R}^{(M+1) \times D}$ indicates the learnable position embedding of the s-th stripe, and $J_s \in \mathbb{R}^{(M+1) \times D}$ is the camera embedding of the s-th stripe. Correspondingly, we utilize Eq. 2 and Eq. 3 to obtain the output of the l-th transformer layer for the s-th stripe $\{Z_s^l \in \mathbb{R}^{(M+1) \times D} | l = 1, 2, \ldots, L\}$.

Finally, we extract the class token cls_s of each stripe from the output of the last intra-part layer Z_s^L, and utilize the ID loss and the triplet loss [9] to optimize the deep model. The ID loss and the triplet loss of the s-th stripe are denoted as $\mathcal{L}_{ID}^s$ and $\mathcal{L}_{TRI}^s$, respectively.

Part-global Layer. In order to learn the completed part interaction, we propose the part-global layer to learn the part-global interaction, as shown in the blue box of Fig. 3. The part-global interaction indicates the interaction between each stripe and all patch tokens from the whole pedestrian image. Here, each stripe is represented by a feature vector, and the interaction between the stripe and all patch tokens in the whole pedestrian image is implemented by the cross-attention mechanism, thus aggregating global information for each stripe. The output of the l-th part-global layer $Y^l \in \mathbb{R}^{S\times D}$ is defined as:

$$Y^l = \mathcal{I}(T^{l-1}, G^{l-1}), \tag{9}$$

where $\mathcal{I}$ is implemented by the multi-head cross-attention (MCA), MLP and LN, and $T^{l-1} \in \mathbb{R}^{S\times D}$ and $G^{l-1} \in \mathbb{R}^{N\times D}$ represent the fused class tokens and the patch tokens from the whole pedestrian image, respectively. In particular, G^{l-1} are obtained by merging the patch tokens of all stripes from $\{Z_s^l | s = 1, 2, \ldots, S\}$. Since Z_s^l learns the interaction between the patch tokens in the s-th stripe, G^{l-1} not only contains the stripe part information, but also represents the global pedestrian representation. Hence, the multi-head cross-attention in $\mathcal{I}$ is implemented by $MCA(T^{l-1}, G^{l-1})$ of the l-th part-global layer, and it is defined as:

$$MCA(T^{l-1}, G^{l-1}) = \text{cat2}(CA_1^l; \cdots; CA_h^l; \cdots; CA_H^l)\hat{U}^l, \tag{10}$$

$$CA_h^l = \beta_h^l G^{l-1} \hat{W}_v^{l,h}, \tag{11}$$

where $\hat{U}^l \in \mathbb{R}^{D\times D}$ and $\hat{W}_v^{l,h} \in \mathbb{R}^{D\times d}$ are the linear projections, and the affinity matrix β_h^l is defined as:

$$\beta_h^l = Softmax(\frac{(T^{l-1}\hat{W}_q^{l,h})(G^{l-1}\hat{W}_k^{l,h})^T}{\sqrt{d}}), \tag{12}$$

where $\hat{W}_q^{l,h} \in \mathbb{R}^{D\times d}$ and $\hat{W}_k^{l,h} \in \mathbb{R}^{D\times d}$ are the linear projections.

Correspondingly, the fused class tokens T^{l-1} in Eq. 9 are defined as:

$$T^{l-1} = \mathcal{F}(Y^{l-1}, C^{l-1}), \tag{13}$$

where $\mathcal{F}$ is implemented by MCA and LN. Similar to Eq. 10, the multi-head cross-attention in $\mathcal{F}$ is implemented by $MCA(Y^{l-1}, C^{l-1})$ of the l-th part-global layer. Here, $C^{l-1} \in \mathbb{R}^{S\times D}$ is the class tokens of all stripes extracted from $\{Z_s^l | s = 1, 2, \ldots, S\}$. $Y^{l-1} \in \mathbb{R}^{S\times D}$ is the output of the $(l-1)$-th part-global layer, and it is initialized to the learnable parameters when $l = 1$. Hence, the fused class tokens could represent the part information more flexibly due to interacting the learnable information.

After L part-global layers, we obtain the output Y^L of the L-th part-global layer and then aggregate it using the max pooling to obtain the global token. Finally, we apply the ID loss $\mathcal{L}_{ID}^g$ and the triplet loss $\mathcal{L}_{TRI}^g$ to optimize the deep model.

C. Adaptive Refined Tokens Module

The number of patch tokens in the transformer network for person ReID is fixed [9], [20], [47], which results in failing to capture effective regional information. To overcome this limitation, we propose the ART module to retain the patch tokens with more information in each stripe during the learning process of the intra-part interaction. The ART module generates the part masks to select the informative patch tokens in each stripe, which is beneficial to learn accurate fine-grained information of human body parts.

In particular, the part transformer layer learns the completed part interaction, and each element in the affinity matrix represents the attention between the patch tokens, that is, the larger the value is, the more information the patch token aggregates. Hence, we design the ART module to generate the part masks based on the affinity matrices of the part transformer layers. It is important to notice that since the part transformer layer consists of the intra-part layer and the part-global layer, the output of the intra-part layer is treated as the input of the part-global layer. The affinity matrix from the part-global layer contains more information of interaction. Hence, we utilize the affinity matrix from the part-global layer as the affinity matrix of the part transformer layer. Correspondingly, the s-th row of the affinity matrix $\beta_h^l \in \mathbb{R}^{S\times N}$ in the l-th part-global layer represents the attention weights of the s-th stripe, and it is denoted as $\beta_{s,h}^l$. Since each head in each layer learns different representations, we sum the attention weights of all heads in previous part-global layers:

$$\hat{\beta}_s^l = \frac{1}{H}\sum_{j=1}^{l-1}\sum_{h=1}^{H}\beta_{s,h}^j[((s-1)\cdot M+1):(s\cdot M)], \tag{14}$$

where $s = 1, 2, \cdots, S$, $l = 2, \cdots, L$, and $\hat{\beta}_s^l \in \mathbb{R}^{1\times M}$. The part mask $\mathcal{M}_s^l \in \mathbb{R}^{1\times M}$ of the s-th stripe is formulated as:

$$\mathcal{M}_s^l[i] = \begin{cases} 1, & \hat{\beta}_s^l[i] > \tau \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

where $i = 1, 2, \cdots, M$, and τ is the threshold to retain the informative patch tokens for each stripe.

Finally, we insert the part masks into the intra-part layer to learn the interaction between the patch tokens with more information, and the corresponding affinity matrix in the l-th intra-part layer is redefined as:

$$\alpha_{s,h}^l = Softmax(\frac{\mathcal{R}(\mathcal{M}_s^l) \odot (Q_{s,h}^l(K_{s,h}^l)^T)}{\sqrt{d}}), \tag{16}$$

where $\mathcal{R}(\mathcal{M}_s^l) \in \mathbb{R}^{M\times M}$ indicates to repeat $\mathcal{M}_s^l$ in the row direction, and $\odot$ denotes the element-wise multiplication. Note that the part mask only selects the patch tokens and abandons the class token for each stripe. In a word, the proposed ART module retains the informative patch tokens in each stripe of the pedestrian image so as to learn the dependencies between them, which improves the discrimination of the pedestrian features.

Algorithm 1: Training procedure

```
Require: The pedestrian image $X \in \mathbb{R}^{H \times W \times C}$,
         initialized backbone, initialized $Y^0$, $\lambda$ for Eq. 8,
         $\tau$ for Eq. 15;
for $i$ in $[1, num\_iters]$ do
    Divide $X$ into $S$ stripes;
    Obtain $\{Z_s^0 \in \mathbb{R}^{(M+1) \times D} | s = 1, 2, \dots, S\}$ by Eq. 8;
    for $l$ in $[1, L-1]$ do
        Obtain $\{Z_s^l \in \mathbb{R}^{(M+1) \times D} | s = 1, 2, \dots, S\}$ by
          Eq. 2 and Eq. 3;
        Obtain $Y^l \in \mathbb{R}^{S \times D}$ by Eq. 9;
    end
    Obtain $\{\mathcal{M}_s^L \in \mathbb{R}^{1 \times M} | s = 1, 2, \dots, S\}$ by Eq. 15;
    Obtain $\{Z_s^L \in \mathbb{R}^{(M+1) \times D} | s = 1, 2, \dots, S\}$ by Eq. 2
      and Eq. 3;
    Obtain $Y^L \in \mathbb{R}^{S \times D}$ by Eq. 9;
    Calculate the final loss $\mathcal{L}$ by Eq. 17;
    Backward to update the deep model;
end
```

D. Optimization

We employ the ID loss and the triplet loss to optimize the deep network, and the overall loss of the proposed CPT is formulated as:

$$\mathcal{L} = \mathcal{L}_{ID}^g + \mathcal{L}_{TRI}^g + \frac{1}{S} \sum_{s=1}^{S} (\mathcal{L}_{ID}^s + \mathcal{L}_{TRI}^s). \tag{17}$$

The training procedure of the proposed CPT is shown in Algorithm 1, where the part masks generated by the proposed ART module are inserted in the last intra-part layer.

IV. Experimental Results

In this section, we first introduce the datasets and the evaluation protocols, and present the implementation details. Then, we conduct the ablation studies to verify the effectiveness of the key components of our method, and compare our method with the state-of-the-art approaches on the four person ReID datasets. Afterwards, we analyze the influence of important parameters for the proposed method. Finally, we visualize our results for intuitive verification.

A. Datasets and Evaluation Protocols

We evaluate the proposed method on four person ReID datasets, i.e., MSMT17 [49], Market-1501 [50], DukeMTMC-reID [51], and CUHK03 [52].

MSMT17 [49] consists of $126,441$ annotated images of $4,101$ identities captured by 15 cameras, and it is a large-scale dataset closer to the real scene because of covering multiple scenes and multiple time periods. MSMT17 contains $32,621$ images of $1,041$ identities for training and $93,820$ images of $3,060$ identities in the test stage.

Market-1501 [50] (Market) consists of $32,668$ annotated images of $1,501$ identities captured by 6 cameras, where each pedestrian is captured by at least 2 cameras, and it contains $12,936$ training images of 751 identities and $19,732$ test images of 750 identities. The pedestrian detection rectangles of 3368 query images are drawn manually, while the pedestrian detection rectangles in the gallery are detected by the deformable part model [53].

DukeMTMC-reID [51] (Duke) has $36,411$ annotated images taken by 8 cameras, in which it contains $16,522$ training images of 702 identities and $19,889$ test images of another 702 identities.

CUHK03 [52] consists of $1,467$ identities captured by 5 cameras, where 767 identities are used for training and the other 700 identities for testing. The dataset contains two kinds of settings which are the labeled images and the detected images. The labeled images consist of $7,368$ training images and $6,728$ testing images, and the detected images contain $7,365$ training images and $7,732$ test images.

The evaluation protocols are the mean average precision (mAP) and the Cumulated Matching Characteristics (CMC) at Rank-1 (R1), Rank-5 (R5) and Rank-10 (R10) accuracies. The post-processing methods are not used for inference, such as re-ranking or multi-query fusion.

B. Implementation Details

In the experiments, we utilize ViT [16] or DeiT [42] as the backbone, and ViT is initialized on ImageNet-21K and then fine-tuned on ImageNet-1K, while DeiT is initialized on ImageNet-1K. Meanwhile, we initialize the class tokens $\{cls_s \in \mathbb{R}^{1 \times D} | s = 1, 2, \dots, S\}$ of the stripes by using the class tokens of ViT or DeiT, and initialize the parameters of the part-global layer by using ViT or DeiT. Moreover, all pedestrian images are resized to 256×128 before feeding into the deep network.

In the training stage, the pedestrian images are augmented by random horizontal flipping, random erasing, random cropping and padding [54]. The batch size is set to 64 which includes 16 identities, and each identity contains 4 pedestrian images. The deep network is optimized by the SGD optimizer with a momentum of 0.9 and the weight decay of 1e-4 [9], [47]. The number of epochs is set to 160, and the learning rate is initialized to 0.01 with the cosine learning rate decay. Unless otherwise specified, the parameter λ in Eq. 8 is set to 3.0, $K = 16$ for each pedestrian image, $L = 12$, $H = 12$, $D = 768$ and $d = 64$ in the part transformer layer, the number of stripes S is set to 2, and the threshold τ in Eq. 16 is set to 0.3.

In the test stage, we concatenate the global token from the part-global layer and the class tokens from the intra-part layer as the final representation.

C. Ablation Studies

In this subsection, we conduct ablation studies to investigate the effectiveness of each component in the proposed method, and their results are listed in Table I. Baseline [9] learns the global interaction with the camera embedding, and it utilizes ViT-B/16 as the backbone. In Table I, IP and PG represent the intra-part layer and the part-global layer, respectively, PP denotes that only learning the interaction between the stripe

TABLE I
THE RESULTS (%) OF ABLATION STUDIES ON MSMT17 AND MARKET. HERE, IP AND PG REPRESENT THE INTRA-PART LAYER AND THE PART-GLOBAL LAYER, RESPECTIVELY, PP DENOTES THAT ONLY LEARNING THE INTERACTION BETWEEN THE STRIPE PARTS, THE SYMBOL § REPRESENTS THAT THE AFFINITY MATRIX OF THE INTRA-PART LAYER IS USED IN THE ART MODULE, AND W/O STRIPES INDICATES THAT IN THE TEST STAGE WE ONLY APPLY THE GLOBAL TOKEN FROM THE PART-GLOBAL LAYER WITHOUT CONCATENATING THE CLASS TOKENS FROM THE INTRA-PART LAYER.

Methods	MSMT17		Market	
	mAP	R1	mAP	R1
Baseline	61.9	81.8	87.7	94.8
IP	64.1	82.3	88.4	95.0
IP+PP	64.8	82.5	88.7	95.2
IP+PG	66.7	83.8	90.3	95.8
Ours (IP+PG+ART)	**68.0**	**84.6**	**91.9**	**96.7**
Ours (IP+PG+ART§)	66.8	84.0	90.5	96.0
Ours (w/o stripes)	67.2	84.1	91.2	96.1

TABLE II
COMPARISON RESULTS (%) WITH THE CNNS-BASED METHODS LEARNING THE PART INFORMATION ON MARKET AND DUKE. TRIPLET DENOTES THE TRIPLET LOSS.

Backbone	Methods	Market		Duke	
		mAP	R1	mAP	R1
CNNs	PCB [2]	77.4	92.3	66.1	81.7
	PCB + Triplet	80.7	93.5	69.5	83.6
DeiT-B/16	IP	87.6	94.5	80.5	89.9
	IP+PG	89.6	95.3	81.9	90.5
	Ours (IP+PG+ART)	**91.0**	**96.0**	**83.2**	**90.8**

parts, the symbol § represents that the affinity matrix of the intra-part layer is used in the ART module, and w/o stripes indicates that in the test stage we only apply the global token from the part-global layer without concatenating the class tokens from the intra-part layer. From the table, we can draw the following conclusions.

Firstly, only learning the intra-part interaction via the proposed intra-part layer (IP) achieves better performance than learning the global interaction (Baseline). Hence, building the interaction within each stripe is beneficial to person ReID.

Secondly, the proposed part-global layer (PG) further enhances IP by 2.6% and 1.9% in terms of mAP on MSMT17 and Market, which demonstrates that the part-global layer could incorporate more information when learning part interaction. Meanwhile, we merge the class token of each stripe after learning the intra-part interaction and build the interaction between the stripe parts of pedestrian image using the cross-attention mechanism. It is denoted as PP. Corresponding, only learning the interaction between the stripe parts (IP+PP) achieves worse performance than learning the interaction between part and global tokens (IP+PG), which demonstrates the effectiveness of the part-global interaction. Furthermore, IP+PG achieves better results than Baseline, which verifies the effectiveness of the designed part transformer layer and it is good at learning the completed part interaction.

Thirdly, the performance improves on the basics of IP+PG when the ART module is introduced. It demonstrates that the part masks generated by the proposed ART module could help remove some irrelevant patch tokens. Meanwhile, the interaction learning of the retained patch tokens is beneficial to improve the discrimination of the pedestrian features. Moreover, using the affinity matrix of the part-global layer as the affinity matrix of the part transformer layer (IP+PG+ART) achieves better results than using the affinity matrix of the intra-part layer (IP+PG+ART§), which verifies the benefits of using the affinity matrix in the part-global layer. It is because the output of the intra-part layer is the input of the part-global layer, and the affinity matrix of the part-global layer contains more interaction information.

Finally, we can see that if only the global token is used in the test stage (training unchanged), the performance (Ours (w/o stripes)) is also impressive, which suggests to only use the global token as an efficient variation with lower storage cost and computational cost in the test stage.

Furthermore, in order to prove the effectiveness of learning part information under the framework of Transformer in the proposed method, we conduct the experiments with the CNNs-based methods learning the part information on Market and Duke. Specifically, we choose the representative CNNs-based method named PCB [2] which also learns the part information via splitting the rigid stripes, and use DeiT-B/16 (ImageNet-1K pre-training) as the backbone for fair comparison. The comparative results are listed in Table II. From the table, we can see that the proposed intra-part layer (IP) only learning the intra-part interaction outperforms PCB by 10.6% and 14.8% in terms of mAP on Market and Duke, respectively. Meanwhile, IP achieves better results than PCB combined with the triplet loss (PCB + Triplet). Hence, the effectiveness of the proposed based on Transformer method learning part information under the framework of Transformer is demonstrated for person ReID.

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

In this subsection, we compare the proposed method with the state-of-the-art methods on four person ReID datasets (MSMT17, Market, Duke, CUHK03), and the results are listed in Table III.

Results on MSMT17. From the table, we can see that using the transformer model of capturing long-range dependencies as the backbone (i.e., the blocks 2 and 3 in Table III) achieves promising results on MSMT17 which is a large-scale dataset. More importantly, the proposed method (Ours† (ViT-B/16)) achieves a new state-of-the-art performance. For example, it obtains 68.0% mAP and 84.6% Rank-1 accuracy, which outperforms NFormer* [47] simultaneously employing CNNs and the transformer model by 5.8% in mAP and 3.8% in Rank-1 accuracy. For a fair comparison with the CNNs-based methods, we use DeiT-B/16 (ImageNet-1K pre-training) as the backbone (Ours† (DeiT-B/16)), and it also achieves promising results. Hence, the effectiveness of the proposed method for person ReID is demonstrated.

Results on Market and Duke. The results are shown in Table III, in which the proposed method achieves comparable

TABLE III
COMPARISON RESULTS (%) WITH THE STATE-OF-THE-ART METHODS FOR PERSON REID ON MSMT17, MARKET, DUKE AND CUHK03. HERE, THE SYMBOL † REPRESENTS THE CAMERA INFORMATION, ∗ INDICATES THAT THEY USE CNNS AS THE BACKBONE AND ALSO EMPLOY THE TRANSFORMER MODEL, AND THE BOLD AND UNDERLINE TEXTS DENOTE THE BEST AND RUNNER-UP RESULTS, RESPECTIVELY.

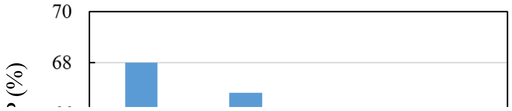
Backbone	Methods	MSMT17		Market		Duke		CUHK03-L		CUHK03-D	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
CNNs	PCB+RPP [2]	40.4	68.2	81.6	93.8	69.2	83.3	-	-	57.5	63.7
	OSNet [55]	52.9	78.7	84.9	94.8	73.5	88.6	-	-	67.8	72.3
	Pyramid [56]	-	-	88.2	95.7	79.0	89.0	76.9	78.9	74.8	78.9
	CBN† [57]	42.9	72.8	77.3	91.3	67.3	82.5	-	-	-	-
	RGA-SC [58]	57.5	80.3	88.4	96.1	-	-	77.4	81.1	74.5	79.6
	ISP [59]	-	-	88.6	95.3	80.0	89.6	74.1	76.5	71.4	75.2
	CDNet [60]	54.7	78.9	86.0	95.1	76.8	88.6	-	-	-	-
	CAL [61]	56.2	79.5	87.0	94.5	76.4	87.2	-	-	-	-
	PAT* [43]	-	-	88.0	95.4	78.2	88.8	-	-	-	-
	HAT* [20]	61.2	82.3	89.5	95.6	81.4	90.4	80.0	82.6	75.5	79.1
	DRL-Net* [45]	55.3	78.4	86.9	94.7	76.6	88.1	-	-	-	-
	NFormer* [47]	62.2	80.8	**93.0**	95.7	**85.7**	<u>90.6</u>	79.1	79.0	76.4	79.0
DeiT-B/16	TransReID† [9]	63.9	82.7	88.0	94.7	81.2	90.1	-	-	-	-
	DCAL [62]	62.3	83.1	87.2	94.5	80.2	89.6	-	-	-	-
	Ours	66.4	<u>84.1</u>	91.0	96.0	83.2	**90.8**	81.0	83.4	77.1	80.4
	(										
ViT-B/16	Trans										
	DC										
	PF										
	(										
	(										

results. It is observed
B/16)) without the can
which learns robust hu
backbone and a transf
3.2% and 4.8% in t
respectively. Furtherm
B/16)) outperforms P
features via a transfor
basic of pose informati
learning the complete
better pedestrian repre

Results on CUHI
both the manually lab
CUHK03. From Table
of the proposed meth
results on the two kind
Hence, the effective
method is verified for

E. Parameters Analys

In this subsection,
important parameters f
of stripes S, the par
part masks in the in
Eq. 15. Note that we
experimental results c
datasets.



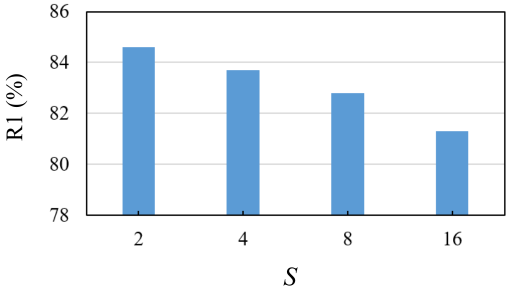


Fig. 5. Evaluation of Rank-1 (R1) accuracy with different number of stripes S on MSMT17.

Number of Stripes S. As shown in Fig. 4 and Fig. 5, we make evaluation of different number of stripes S. From the

EVALUATION OF PI

d	
	m
32	6
64	**6**
128	6
256	6
512	6

EVALUATION OF PERFOR
INTRA-PART LAYER
ACCUMULATION OPERAT

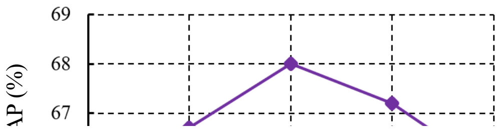
Layer	
	mAP
2	55.5 (55
3	55.3 (54
4	57.1 (58
5	59.9 (59
6	60.8 (59
7	62.6 (61
8	64.5 (62
9	66.1 (64
10	66.7 (65
11	67.3 (66
12	**68.0** (67

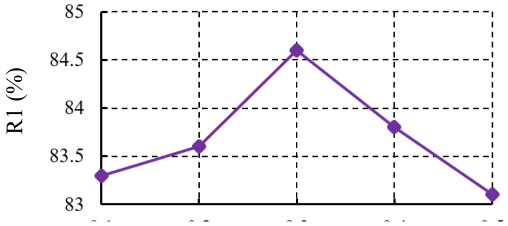
figure, we can see that
as S gets bigger, the r
when we divide more stripes, less patch t
to learn the interaction, which leads to in
for completed part interaction. Hence, we
default setting.

Parameter d**.** As shown in Table IV
parameter d to explore its effect for the perf
ReID. Here, we vary d from 32 to 512. F
can see that the best performance for perso
at 64 for the parameter d. Hence, we set d =
setting.

Inserted Location of the Part Masks. We analyze the inserted location of the part masks generated by the ART module in the intra-part layer, and the experimental results are listed in Table V. From the table, we can see that it is performed best in the last intra-part layer. It is because the part masks are obtained by summing the affinity matrices of all previous part-global layers, and therefore the part masks inserted in the last part-global layer contain more interaction information. Furthermore, we also conduct the experiments without using the accumulation operation in Eq. 14, and the evaluation results are listed in the brackets. From the table, we can see that the performance gets worse when the part masks are not obtained by summing the affinity matrices of all previous part-global layers.

Threshold τ**.** We perform the evaluation experiments with different values of threshold τ in Eq. 15, and the results are shown in Fig. 6 and Fig. 7. Here, τ controls how many patch tokens in each stripe are retained to conduct the interaction. From the figure, we can see that the performance degrades when $\tau < 0.3$, because too many patch tokens may introduce irrelevant information. Meanwhile, the performance drops when $\tau > 0.3$, because a small number of patch tokens is not enough to learn sufficient interaction between the patch tokens. Hence, we set $\tau = 0.3$.





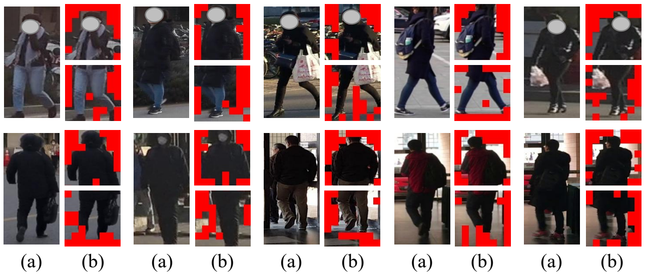


Fig. 8. (a) Input pedestrian images, (b) visualization results of the part masks generated by the proposed ART module, where the red blocks indicate the discarded patch tokens.

F. Visualization

We show the visualization results of the part masks generated by the proposed ART module in Fig. 8, where we randomly select ten pedestrian images with different identities from MSMT17. Here, the red blocks indicate the discarded patch tokens. From the figure, we can see that the part masks could retain the patch tokens with more information in each stripe and then we learn the interaction between them, which is beneficial to capture fine-grained information of human body parts, so that improving the discrimination of the pedestrian representations.

V. Conclusion

In this paper, we have proposed CPT to learn the completed part interaction via the well-designed part transformer layer for person ReID, where the part transformer layer could learn long-range dependencies from the aspects of the intra-part interaction and the part-global interaction. Furthermore, we propose the ART module to retain the informative patch tokens in each pedestrian image, where we utilize the part masks to establish the dependencies of the patch tokens with more information in each stripe of pedestrian image so as to improve the discrimination of the pedestrian features. The experimental results on four person ReID datasets have demonstrated the effectiveness of the proposed method.

References

[1] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 384–393.

[2] 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 *Proceedings of the European Conference on Computer Vision*, 2018, pp. 480–496.

[3] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.

[4] F. Xu, B. Ma, H. Chang, and S. Shan, "Prdp: Person reidentification with dirty and poor data," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 11 014–11 026, 2022.

[5] D. Cheng, Z. Li, Y. Gong, and D. Zhang, "Fusion of multiple person re-id methods with model and data-aware abilities," *IEEE Transactions on Cybernetics*, vol. 50, no. 2, pp. 561–571, 2020.

[6] Z. Zhang, Y. Wang, S. Liu, B. Xiao, and T. S. Durrani, "Cross-domain person re-identification using heterogeneous convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1160–1171, 2022.

[7] B. Sun, Y. Ren, and X. Lu, "Semisupervised consistent projection metric learning for person reidentification," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 738–747, 2022.

[8] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019.

[9] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 15 013–15 022.

[10] T. Si, F. He, Z. Zhang, and Y. Duan, "Hybrid contrastive learning for unsupervised person re-identification," *IEEE Transactions on Multimedia*, 2022, doi:10.1109/TMM.2022.3174414.

[11] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, "Pose-guided representation learning for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 622–635, 2022.

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

[13] Z. Zhang, H. Zhang, and S. Liu, "Person re-identification using heterogeneous local graph attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 136–12 145.

[14] Q. Zhang, J. Lai, Z. Feng, and X. Xie, "Seeing like a human: Asynchronous learning with dynamic progressive refinement for person re-identification," *IEEE Transactions on Image Processing*, vol. 31, pp. 352–365, 2022.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[16] 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," 2020, arXiv preprint arXiv:2010.11929.

[17] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[18] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8126–8135.

[19] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two pure transformers can make one strong gan, and that can scale up," in *Proceedings of the Advances in Neural Information Processing Systems*, 2021, pp. 14 745–14 758.

[20] G. Zhang, P. Zhang, J. Qi, and H. Lu, "Hat: Hierarchical aggregation transformers for person re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 516–525.

[21] S. Liao and L. Shao, "Transmatcher: Deep image matching through transformers for generalizable person re-identification," in *Proceedings of the Advances in Neural Information Processing Systems*, 2021, pp. 1992–2003.

[22] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 536–551.

[23] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[24] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[25] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2012.

[26] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3685–3693.

[27] Z. Wang, R. Hu, C. Chen, Y. Yu, J. Jiang, C. Liang, and S. Satoh, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 3006–3020, 2018.

[28] Y. Feng, Y. Yuan, and X. Lu, "Person reidentification via unsupervised cross-view metric learning," *IEEE Transactions on Cybernetics*, vol. 51, no. 4, pp. 1849–1859, 2021.

[29] Y. Li, T. Zhang, L. Duan, and C. Xu, "A unified generative adversarial framework for image generation and person re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2018, pp. 163–172.

[30] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1288–1296.

[31] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 3988–3994.

[32] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, pp. 1–20, 2018.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1487–1495.

[34] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[35] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," 2017, arXiv preprint arXiv:1711.08184.

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

[37] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, 2019.

[38] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 365–381.

[39] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9637–9646.

[40] X. Ren, D. Zhang, X. Bao, and Y. Zhang, "S^2-Net: Semantic and salient attention network for person re-identification," *IEEE Transactions on Multimedia*, 2022, doi: 10.1109/TMM.2022.3174768.

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 213–229.

[42] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 10 347–10 357.

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2898–2907.

[44] Z. Ma, Y. Zhao, and J. Li, "Pose-guided inter-and intra-part relational transformer for occluded person re-identification," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 1487–1496.

[45] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Transactions on Multimedia*, 2022, doi:10.1109/TMM.2022.3141267.

[46] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2540–2549.

[47] H. Wang, J. Shen, Y. Liu, Y. Gao, and E. Gavves, "NFormer: Robust person re-identification with neighbor transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7297–7307.

[48] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "ABD-Net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8350–8360.

[49] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.

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

[51] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 17–35.

[52] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[53] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

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

[55] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3701–3711.

[56] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8506–8514.

[57] Z. Zhuang, L. Wei, L. Xie, T. Zhang, H. Zhang, H. Wu, H. Ai, and Q. Tian, "Rethinking the distribution gap of person re-identification with camera-based batch normalization," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 140–157.

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

[59] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 346–363.

[60] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6725–6734.

[61] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1005–1014.

[62] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4692–4702.

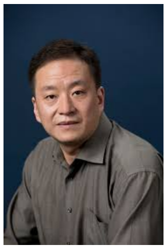
Zhong Zhang (Senior Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He is a Professor with Tianjin Normal University, Tianjin, China. He has published about 120 articles in international journals and conferences such as the IEEE Transactions on Fuzzy Systems, IEEE Transactions on Multimedia, Pattern Recognition, IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Circuits Systems Video Technology, IEEE Transactions on Information Forensics and Security, Signal Processing (Elsevier), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and AAAI Conference on Artificial Intelligence (AAAI). His research interests include computer vision, and deep learning.

Di He is a master student at Tianjin Normal University, Tianjin, China. His research interests include person re-identification and deep learning.

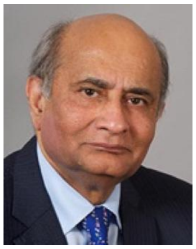
Shuang Liu (Senior Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014.

She is a Professor with Tianjin Normal University, Tianjin, China. She has published over 60 articles in major international journals and conferences. Her research interests include computer vision, and deep learning.



Baihua Xiao received the B.S. degree in Electronic Engineering from Northwestern Polytechnical University, Xian, China and the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1995 and 2000, respectively.

From 2005, he has been a Professor at the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, computer vision, image processing and machine learning.



Tariq S. Durrani is Research Professor at University of Strathclyde, Glasgow Scotland. His research covers AI, Signal Processing and Technology Management. He has authored 400 publications; supervised 45 PhDs.

He is a Fellow of the: IEEE, UK Royal Academy of Engineering, Royal Society of Edinburgh, IET, and The Third World Academy of Sciences. He was elected Foreign Member of the Chinese Academy of Sciences and the US National Academy of Engineering in 2021 and 2018, respectively.