Occlusion-Aware Transformer With Second-Order Attention for Person Re-Identification

Yanping Li, Yizhang Liu, Hongyun Zhang, Cairong Zhao, *Member, IEEE*, Zhihua Wei, and Duoqian Miao

***Abstract*— Person re-identification (ReID) typically encounters varying degrees of occlusion in real-world scenarios. While previous methods have addressed this using handcrafted partitions or external cues, they often compromise semantic information or increase network complexity. In this paper, we propose a new method from a novel perspective, termed as OAT. Specifically, we first use a Transformer backbone with multiple class tokens for diverse pedestrian feature learning. Given that the self-attention mechanism in the Transformer solely focuses on low-level feature correlations, neglecting higher-order relations among different body parts or regions. Thus, we propose the Second-Order Attention (SOA) module to capture more comprehensive features. To address computational efficiency, we further derive approximation formulations for implementing second-order attention. Observing that the importance of semantics associated with different class tokens varies due to the uncertainty of the location and size of occlusion, we propose the Entropy Guided Fusion (EGF) module for multiple class tokens. By conducting uncertainty analysis on each class token, higher weights are assigned to those with lower information entropy, while lower weights are assigned to class tokens with higher entropy. The dynamic weight adjustment can mitigate the impact of occlusion-induced uncertainty on feature learning, thereby facilitating the acquisition of discriminative class token representations. Extensive experiments have been conducted on occluded and holistic person re-identification datasets, which demonstrate the effectiveness of our proposed method.**

***Index Terms*— Second-order attention, information entropy, uncertainty.**

I. Introduction

PERSON Re-identification (ReID) aims to locate and retrieve the same pedestrian across multiple cameras in complex environments [1], [2], [3], which has been widely used in applications such as security monitoring, criminal investigation, missing persons search, and intelligent traffic management [4], [5]. In real-world scenarios, occlusion is often the case, as illustrated in Fig. 1. Pedestrians may be obstructed by objects such as trees, vehicles, billboards, and other pedestrians. The diversity in terms of size, shape, color, and position of occlusions presents a significant challenge for general holistic person ReID methods in obtaining discriminative pedestrian features [6], [7].

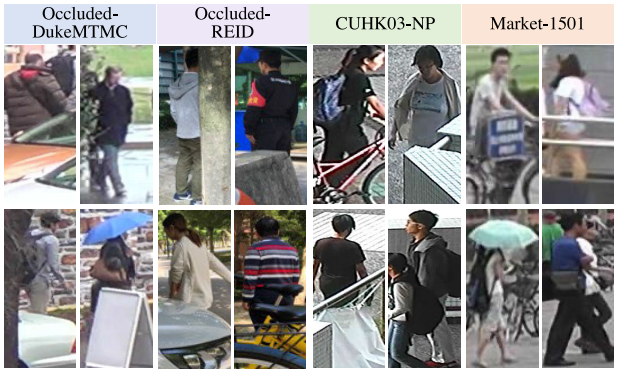


Fig. 1. Examples of occluded person images from two occluded ReID datasets and two holistic ReID datasets.

In recent years, many researchers have been delving into methods for occluded person ReID, with partitioning-based [8], [9], [10] and external cues-based [11], [12], [13], [14], [15] methods emerging as two popular solutions. Although these methods work well by learning well-aligned part features, they each also come with their inherent drawbacks. Partitioning pedestrian images or feature maps can to some extent alleviate the impact of occlusion, but it inevitably impairs the learning of contextual information among adjacent parts, hindering the integrity of local semantic information. External cues typically encompass pose estimation and human parsing, both of which contribute to the identification of occlusions. However, the incorporation of external models increases the overall model complexity, demanding more training time and computational resources. In addition, within intricate occlusion scenarios, external cues-based methods are susceptible to noise interference, resulting in false detections.

In this paper, we propose a novel approach to addressing the occlusion problem that is designed to operate independently of handcrafted partitioning and external cues. Specifically, the Transformer is used as the backbone of the proposed method due to its superior long-range modeling ability [16], [17], [18]. Given that the self-attention module in the Transformer primarily emphasizes low-level feature correlations,

Manuscript received 10 October 2023; revised 1 March 2024; accepted 21 April 2024. Date of publication 30 April 2024; date of current version 6 May 2024. This work was supported in part by the National Key Research and Development Program under Grant 2022YFB3104700; and in part by the National Natural Science Foundation of China under Grant 61976158, Grant 62376198, Grant 62076182, Grant 62163016, and Grant 62006172. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yi Yang. *(Corresponding author: Duoqian Miao.)*

Yanping Li, Hongyun Zhang, Cairong Zhao, Zhihua Wei, and Duoqian Miao are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: liyp8023yz@tongji.edu.cn; zhanghongyun@tongji.edu.cn; zhaocairong@tongji.edu.cn; zhihua_wei@tongji.edu.cn; dqmiao@tongji.edu.cn).

Yizhang Liu is with the School of Software Engineering, Tongji University, Shanghai 201804, China (e-mail: lyz8023lyp@gmail.com).

Digital Object Identifier 10.1109/TIP.2024.3393360

it inherently restricts higher-order relations among different body parts or regions, which are particularly crucial for occluded person Re-ID [1], [19]. To address this issue, we propose the Second-Order Attention (SOA) module to delve into contextual information within attention weight. This module utilizes the Laplace matrix in spectral clustering to extract features and obtains the contextual information of the attention weights using a series of differentiable operations. Notably, the attention weights emphasize specific areas of pedestrian images that may be associated with key features of the pedestrians. The context of the attention weights highlights regions or non-occluded parts with salient relations across the image, which helps the model to better understand and differentiate between pedestrians. Meanwhile, considering computational efficiency, we derive an approximate implementation for SOA, reducing the computational complexity from cubic complexity to linear complexity. To fully exploit the contextual information of both self-attention and the proposed second-order attention mechanisms, we fuse them using a summation operation, resulting in a more comprehensive and discriminative feature representation. Additionally, to render the model occlusion-aware, we introduce multiple class tokens to extract pedestrian features from various subspaces (with distinct semantic information). Observing that some class tokens learn features of the visible part of the pedestrian, while others learn the features of occlusions, we propose the Entropy Guided Fusion (EGF) module to alleviate uncertainties caused by occlusions. Higher entropy values indicate less valuable information from class tokens and vice versa. Multiple class tokens are fused with the guidance of information entropy, i.e., higher weights are assigned to those with lower information entropy, while lower weights are assigned to class tokens with higher entropy, thus mitigating the issues caused by occlusions.

The main contributions of our work are as follows:

1) The proposed OAT offers a novel perspective on addressing the occlusion challenge in person Re-ID. It is designed to function autonomously, without relying on handcrafted partitioning or external cues. Furthermore, it demonstrates robustness in handling occluded pedestrians.

2) Two novel modules have been devised: SOA is designed to learn higher-order relations among different body parts or regions, while EGF is proposed to mitigate the uncertainties induced by occlusions. With two modules complementing each other, discriminative pedestrian features can be obtained.

3) Extensive experiments are conducted on both occluded and holistic datasets, including Occluded-DukeMTMC [20], Occluded-REID [21], Market-1501 [22], DukeMTMC-reID [23], CUHK03-NP [24]. The superiority of the proposed method is verified through ablation experiments and visualization comparisons.

II. Related Work

A. Holistic Person Re-Identification

In recent years, significant progress has been achieved in holistic person re-identification using deep learning in conjunction with several well-designed modules, resulting in a substantial enhancement of recognition performance. For example, Zheng et al. [25] introduce the ID-discriminative Embedding (IDE), regarding person ReID as both a classification and verification task, while simultaneously learning discriminative embeddings and similarity measurement. Luo et al. [2] establish a robust baseline by integrating a bag of tricks to enhance discriminative global feature learning. Building upon this foundation, Ye et al. [4] propose the AGW model which includes Non-local Attention (Att) Block, Generalized-mean (GeM) Pooling, and Weighted Regularization Triplet (WRT) loss to augment pedestrian feature learning. Moreover, Zhang et al. [26] cast person ReID as a multi-instance learning problem to achieve strong spatial misalignment tolerance and enhanced discriminative capabilities. Quan et al. [27] propose the Auto-ReID method, an automated neural architecture search approach tailored for ReID tasks, which integrates body structure information into the search space by devising a typical classification search space and a part-aware module. Wu et al. [28] propose a progressive training framework for estimating pseudo-labels of unlabeled data, and updating the CNN model by joint training labeled data, pseudo-labeled data, and index-labeled data. Similarly, Quan et al. [29] employ a progressive learning approach to gradually select unlabeled data with reliable pseudo-labels, thereby enriching the training data and updating the model to narrow the domain gap. Lin et al. [30] propose a cross-camera unsupervised method that iteratively optimizes CNNs and individual sample relationships to address person ReID, effectively overcoming challenges of camera variances and identity similarities.

Several methods propose segmenting pedestrian images to acquire the local features of pedestrians. For instance, PCB [8] generates multiple local features of pedestrians using a simple uniform horizontal partitioning strategy, and has been widely recognized as a strong baseline for learning local features of pedestrians [9], [31], [32], [33], [34]. MGN [31] partitions different numbers of horizontal stripes across distinct branches to acquire multi-granularity local features.

To capture features at different scales and granularities, numerous studies have introduced pyramid structures for person ReID, such as the coarse-to-fine pyramid model proposed by [32] and the striped pyramidal pooling block developed by [35]. Additionally, Chen et al. [36] propose an attention pyramid structure, which enables the model to focus on salient cues at different scales like human visual perception. RGA-SC [37] incorporates relation-aware global attention, stacking relations, and combining shallow convolutional models for better fusion of global and local features. Integrating external cues as the prior knowledge is also a research hotspot. Zhang et al. [38] utilize the trained DensePose [39] to segment the human body into 24 densely semantic regions. Zhu et al. [11] propose Identity-guided Semantic Parsing (ISP), identifying body parts and personal belongs at the pixel level to address misalignment issues.

Recently, investigations have delved into the application of Vision Transformers [16] in person ReID tasks. TransReID [17] pioneers Transformer-based person ReID research, with a specially designed Jigsaw Patch Module (JPM) and Side Information Embedding (SIE) to enhance robust feature

learning. AAformer [40] introduces alignment techniques within the Transformer architecture, enabling automatic localization of human and non-human parts, coupled with joint learning of part alignment and feature representation. HAT [41] leverages the strengths of CNN and Transformers, effectively mining and merging detailed and semantic information from cross-level features to reinforce multi-scale feature extraction.

While these methods contribute to enhancing feature learning of holistic pedestrians, their performance tends to decline when addressing the prevalent issue of occlusion in real-world scenarios. In such situations, various body parts of pedestrians are often obstructed by objects or other pedestrians, which can undermine the effectiveness of the aforementioned approaches.

B. Occluded Person Re-Identification

Occlusion typically results in challenges such as spatial misalignment and incomplete body information. Currently, occluded person Re-ID methods can be broadly categorized into three types: CNN-based methods, external cues-based methods, and Transformer-based methods.

CNN-based methods predominantly concentrate on specific designs built on well-performed architectures like ResNet50. Zhuo et al. [21] systematically define the occlusion problem in person Re-ID, creating the Attentive Focus on Pedestrian Body (AFPB) framework that includes an occlusion simulator and a multi-task loss. FPR [42] employs a foreground probability generator to linearly reconstruct probe spatial features by gallery spatial features for facilitating alignment-free matching. Zhuo et al. [43] devise a Teacher-Student learning framework to alleviate the deficiency of occluded pedestrian images. IGOAS [44] combines batch-based incremental generative occlusion block with a global-adversarial suppression framework to extract global features and non-occluded body features. QPM [45] jointly learns part features and part quality predictions, identifying occlusion scenarios via identity-aware spatial attention, and generating global features of pedestrians by an adaptive global feature extraction module. PRE-Net [10] suppresses occlusion noise and enhances the features of visible parts through various part partition strategies, partial relationship aggregation, and inter-part omnibearing fusion modules. RTGAT [46] jointly reasons about the visible parts of the human body and compensates for the semantic loss of the occluded parts to learn complete human representations in occluded images. Despite their proficiency in capturing visible part features, the performance of CNN-based methods is typically inferior to Transformer-based approaches due to limited long-range modeling dependencies.

External cues-based methods primarily involve integrating pose estimation or human parsing information into CNN or Transformer networks. PGFA [20] utilizes pose landmarks to generate attention maps, divides the global feature into parts, and uses shared-region features for matching. Building upon this foundation, Miao et al. [12] introduce a Pose-Embedded Feature Branch (PEFB) for adaptively adjusting channel-wise feature responses to enhance the learned features. HOReID [47] leverages CNN and key-point estimation to extract semantic information of local features, employs adaptive directional graph convolution layers for relationship propagation, and introduces a cross-graph embedded alignment layer for local feature alignment and similarity prediction. PVPM [7] jointly learns discriminative features with pose-guided attention, while automatically mining part visibility. Yang et al. [48] discretize pose information into body part visibility labels to suppress the impact of occluded regions. Xu et al. [15] propose the Feature Recovery Transformer (FRT) to solve the visible graph matching and feature recovery problem. Somers et al. [49] combine identity with human parsing labels to design a body part attention module to address the occlusion problem. Dou et al. [14] propose the Human Co-parsing Guided Alignment (HCGA) framework for weakly supervised training under the constraints of local spatial consistency, semantic consistency, and background. MSDPA [50] introduces multi-source semantic cues into Transformer, exploring semantic correlations between body parts using a dynamic attention mechanism. Ma et al. [51] present a pose-guided intra- and inter-part relational transformer, improving feature extraction techniques and bolstering inter-part relationship learning. Although these methods enhance the discriminative capability of the model, however, the use of these external cues often relies on additional models, resulting in high complexity and computational resource consumption.

Transformer-based approaches perform well in modeling long-range dependencies of pedestrian features. PAT [52] reidentifies occluded persons by discovering distinct parts and introduces effective learning mechanisms to better capture part prototypes solely with identity labels. DRL-Net [53] employs local features for global reasoning and integrates contrastive feature learning techniques along with data augmentation strategies to mitigate occlusion interference. Mao et al. [54] develop Attention Map Guided (AMG) transformer pruning, which weakens redundant attention heads and parameters through entropy calculation of key dimensions and token importance estimation. Compared to existing CNN-based methods, Transformer-based approaches exhibit stronger robustness against occlusions. In this paper, building upon the Transformer framework, we propose an occlusion-aware Transformer method without relying on external cues. To acquire more distinctive pedestrian features, we introduce the second-order attention module and the entropy guided fusion module to enhance the higher-order relations between different body parts and alleviate the impact of occlusions, thereby yielding richer and more discriminative pedestrian features.

III. Method

In this section, we first describe the overall network architecture of the proposed method. Subsequently, we delve into the details of two novel modules: the Second-Order Attention Module and the Entropy Guided Fusion Module. Lastly, we present the objective function of the model.

A. Network Architecture

The overall framework of the proposed method is depicted in Fig. 2, where ViT [16] is adopted as the backbone. Given

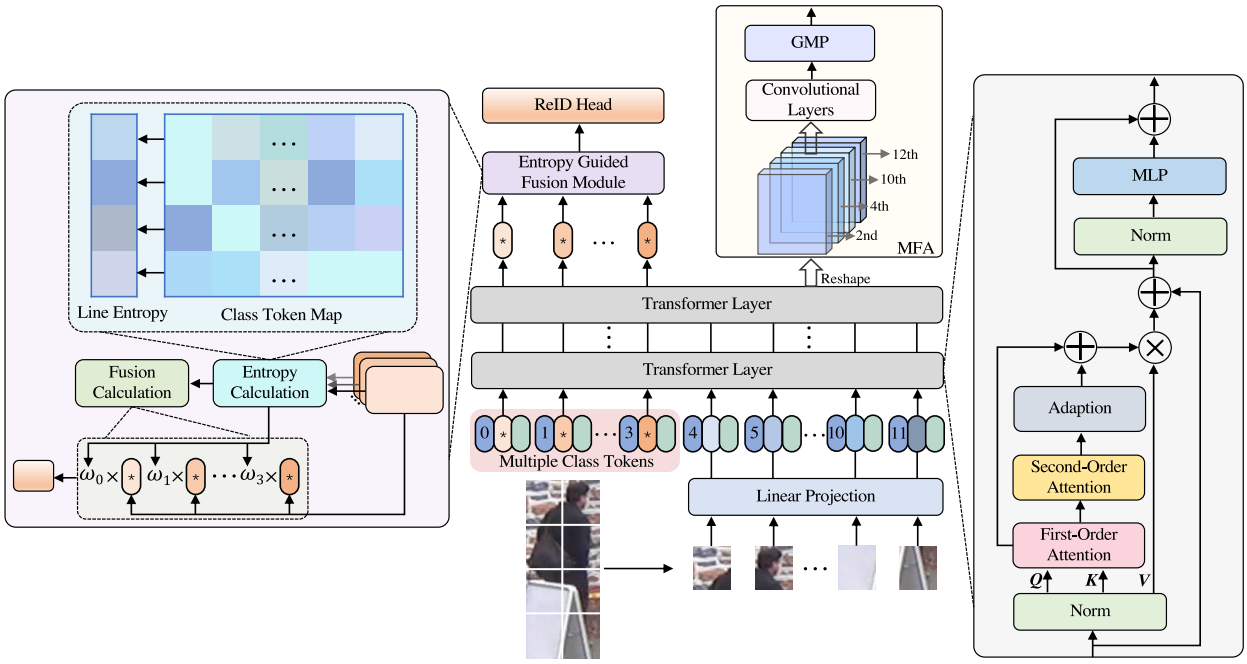


Fig. 2. The architecture of the proposed OAT framework. Specifically, the input sequence is combined with multiple class tokens and jointly fed into the encoder of the Transformer, along with learnable position embeddings and camera embeddings. Then, in each Transformer Layer, a Second-Order Attention module is introduced based on the First-Order Attention (i.e., self-attention) to capture contextual information of attention weight. The output of the Second-Order Attention and the First-Order Attention are merged to facilitate comprehensive feature learning. In the last Transformer Layer, the features learned from multiple class tokens are fused by the Entropy Guided Fusion Module, making the fused features robust to occlusions. Moreover, MFA denotes Multilayer Feature Aggregation that aggregates features from the 2nd, 4th, 10th, and 12th Transformer Layers through convolutional operations. These features constitute what remains after the class tokens have been excluded from the original features.

an input image $X \in \mathbb{R}^{H\times W\times C}$, where H, W, and C are the height, width, and channel dimensions of the image, respectively, we split the image X into N_0 patches $\{x_i|i = 1, 2, \ldots N_0\}$, and each patch is embedded to R dimensions by the linear projection function $\mathcal{F}(\cdot)$. Multiple class tokens are attached to the patch embedding to capture diverse semantic information about pedestrians. In addition, the learnable position embedding $\mathcal{P} \in \mathbb{R}^{(N_0+M)\times R}$ and the camera embedding $\mathcal{C} \in \mathbb{R}^{(N_0+M)\times R}$ are added to the patch embedding for preserving the spatial and camera information of the image [17]. The aforementioned process can be formulated as:

$$\mathcal{Z}_0 = [x^0_{cls}; x^1_{cls}; \ldots; x^{M-1}_{cls}; \mathcal{F}(x_1); \mathcal{F}(x_2); \ldots; \mathcal{F}(x_{N_0})] + \mathcal{P} + \lambda\mathcal{C}, \quad (1)$$

where $\mathcal{Z}_0$ is the input of the first Transformer Layer; M denotes the number of class tokens; λ is a hyperparameter used to trade-off the weight of camera embedding. For simplicity, we will use N to denote $N_0 + M$ in the following.

Given that the self-attention mechanism in the Transformer Layers predominantly highlights low-level feature correlations, limiting the exploration of higher-order relations among distinct body parts or regions. To address this issue, in each Transformer Layer, we introduce the Second-Order Attention Module and integrate it with the First-Order Attention (i.e., self-attention) module to generate attentive features in a hierarchical approach. In other words, the First-Order Attention mainly focuses on low-level feature correlations, while the Second-Order Attention primarily emphasizes higher-order part or region relations. These two attention mechanisms complement each other, resulting in more comprehensive pedestrian features. In addition, to highlight the class tokens corresponding to the unobstructed human parts, we introduce the Entropy Guided Fusion Module in the last Transformer Layer. This module dynamically fuses class tokens based on calculated entropy value and it can mitigate the impact of occlusion-induced uncertainty on feature learning, facilitating the acquisition of discriminative class token representations and enhancing occlusion-aware capabilities.

Meanwhile, to prevent the features in each Transformer Layer from being smoothed and becoming similar, we also utilize a multilayer feature aggregation approach. Specifically, the features at different Transformer Layers are aggregated together to fully exploit the contextual information across hierarchies to capture diverse semantic information. Similar to [55], [56], we aggregate the features from the 2nd, 4th, 10th, and 12th Transformer Layers through convolutional operations. The above process can be formulated as:

$$F_{agg} = \sigma(pool(\psi((f_2, f_4, f_{10}, f_{12})))), \quad (2)$$

where $f_l = Reshape([x^l_1; x^l_2; \ldots; x^l_{N_0}])$, $l \in \{2, 4, 10, 12\}$; ψ denotes the spatial convolutional layer; $pool(\cdot)$ and σ denote the max pooling operation and the $sigmoid$ function, respectively.

The fused class tokens and the features obtained after MFA are utilized for the calculation of the loss function, which supervises the network training. Further details are available in Section III-D. Next, we provide an in-depth explanation of the key components of the proposed method, including

the Second-Order Attention Module in Section III-B and the Entropy Guided Fusion Module in Section III-C.

B. Second-Order Attention Module

The input sequence $\mathcal{Z}_0$ is first mapped into Q, K, $V \in \mathbb{R}^{N\times d}$ using different linear transformations, where d denotes the embedding dimension. The first-order attention weights that imply low-level feature correlations can be calculated as:

$$A = softmax(\frac{Q \times K^T}{\sqrt{d}}), A \in \mathbb{R}^{N\times N}, \tag{3}$$

where $\frac{1}{\sqrt{d}}$ is the scaling factor, and T denotes the transpose operation.

To learn the higher-order relations between different body parts or regions, we construct a global graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here, the nodes $\mathcal{V}$ encompass attention weights of all class tokens and patches, while the edges $\mathcal{E}$ connect all these attention weights. The second-order attention module uses the Laplace matrix in spectral clustering to obtain the contextual information of the attention weights using a series of differentiable operations. Specifically, the second-order attention weights can be represented as:

$$W = A^T A, W \in \mathbb{R}^{N\times N}, \tag{4}$$

where $w_{ij} = \sum_k a_{ki}a_{kj}$ denotes the similarity of the attention weights corresponding to the $i\text{-}th$ and $j\text{-}th$ class tokens or patches. Therefore, the Laplacian matrix can be formulated as follows:

$$L = D - W, \tag{5}$$

where D denotes the diagonal degree matrix of W. The straightforward approach to obtain prominent correlations among attention weights with regard to different class tokens or patches is to perform singular value decomposition (SVD) on L. However, the SVD suffers from non-differentiable and time-consuming drawbacks in network training. Thus, we propose an approximate and differentiable solution to solve this problem.

Proposition 1: Let l_i represent the $i\text{-}th$ column vector of L, The contextual information of the $i\text{-}th$ attention weight can be expressed as:

$$e_i = \sqrt{l_i^T l_i} = \sqrt{(\sum_{j,j\neq i} w_{ij})^2 + \sum_{j,j\neq i} w_{ij}^2} \tag{6}$$

$$\approx \sqrt{2}(\sum_k a_{ki} - \frac{1}{N}(\sum_k a_{ki})^2). \tag{7}$$

Proof: According to Cauchy's Inequality, as $w_{ij} > 0$, the following inequality holds:

$$\frac{1}{N}(\sum_{j,j\neq i} w_{ij})^2 \leq \sum_{j,j\neq i} w_{ij}^2 \leq (\sum_{j,j\neq i} w_{ij})^2. \tag{8}$$

Thus, the upper and lower bounds of e_i can be determined by substituting Eq. 8 into Eq. 6:

$$\sqrt{\frac{N+1}{N}} \sum_{j,j\neq i} w_{ij} \leq e_i \leq \sqrt{2} \sum_{j,j\neq i} w_{ij}. \tag{9}$$

For simplicity, we utilized the upper bound of e_i as an approximate formulation to capture contextual information of the attention weights:

$$\begin{aligned} e_i &= \sqrt{2} \sum_{j,j\neq i} w_{ij} \\ &= \sqrt{2} \sum_j w_{ij} - \sum_k a_{ki}^2 \\ &= \sqrt{2}(\sum_k \sum_j a_{ki}a_{kj} - \sum_k a_{ki}^2) \\ &\approx \sqrt{2}(\sum_k \sum_j a_{ki}a_{kj} - \frac{1}{N}(\sum_k a_{ki})^2). \end{aligned} \tag{10}$$

Since $softmax()$ in Eq. 3 leads to $\sum_j a_{kj} = 1$, we obtain the approximate solution of e_i in Eq. 7.

Accordingly, we can obtain the second-order attention weights $E = \{e_i | i = 1, 2, \ldots, N\}$. After adjusting the dimensionality of the second-order attention weights, we add them to the first-order attention weights, resulting in discriminative pedestrian features from the perspective of low-level correlations and higher-order relations, respectively:

$$F_{att} = (A + Adaption(sigmoid(E))) \times V. \tag{11}$$

C. Entropy Guided Fusion Module

Generally, the class token serves as a global information aggregator. Incorporating multiple class tokens is advantageous for capturing diverse semantic information related to pedestrians. However, simply aggregating the features of multiple class tokens may lead to the accumulation of redundant or similar information, thereby limiting the discriminative feature learning. In real-world scenarios, the diversity in size, shape, and position of occlusions brings great uncertainty, which also restricts the feature learning of the unobstructed human parts. To solve this problem, we introduce information entropy to weaken the uncertainty caused by occlusion. Specifically, given a dataset containing U classes, we denote the class tokens of the last Transformer Layer as $\{y_{cls}^i | i = 0, 1, \ldots, M-1\}$, with $y_{cls}^i \in \mathbb{R}^{1\times R}$, and y_{cls}^i is then mapped to a $1 \times U$ vector through linear transform $\phi(\cdot)$. Following this, the softmax function is applied to compute the probability distribution, denoted as y_i, indicating the likelihood of the class token y_{cls}^i belonging to U classes. The above process can be formulated as:

$$y_i = softmax(\phi(y_{cls}^i)). \tag{12}$$

Consequently, the information entropy value of the class token y_{cls}^i can be determined by:

$$s_i = -\sum_j^U y_i^j log(y_i^j), \tag{13}$$

where y_i^j denotes the j-th element in y_i.

Accordingly, we obtain the overall information entropy $S = \{s_i | i = 0, 1, \ldots, M-1\}$. We argue that a higher information entropy value of a class token indicates relatively limited discriminative and deterministic pedestrian features

have been learned, and vice versa. To obtain class tokens robust to occlusions, we propose the entropy guided fusion module. Specifically, the information entropy matrix S is first normalized as follows:

$$S_{norm} = \frac{S - min(S)}{max(S) - min(S)}. \tag{14}$$

Subsequently, the weights of different class tokens are determined based on the information entropy value. Class tokens with smaller information entropy values should receive larger weights, whereas those with larger entropy values should be assigned smaller weights. The weights for each class token can be computed as:

$$S_{weight} = \frac{1 - S_{norm}}{\sum S_{norm}}. \tag{15}$$

Finally, the class tokens are fused by weighted summation:

$$F_{egf} = \sum_{i}^{M-1} y_{cls}^{i} \cdot S_{weight}^{i}, \tag{16}$$

where S_{weight}^{i} denotes the i-th element in S_{weight}. The entropy guided fusion module effectively balances and integrates features learned by each class token. By maintaining feature diversity and richness while avoiding unnecessary feature redundancy, it produces more comprehensive and complementary pedestrian features.

D. Objective Function

Person Re-ID is treated as a classification task, thus the cross-entropy loss is employed for supervised training:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{U} p_i \log(\hat{p}_i), \tag{17}$$

where $\hat{p}_i$ is the predicted label, while p_i represents the ground truth label. To decrease the intra-class distance while increasing the inter-class distance, we use a triplet loss with soft-margin function:

$$\mathcal{L}_{tri} = \log[1 + \exp(\|f_a - f_p\|_2^2 - \|f_a - f_n\|_2^2)], \tag{18}$$

where f_a refers to the anchor sample; f_p is the positive sample with the same identity as the anchor; f_n is the negative sample with a different identity from the anchor, and $\|\cdot\|_2$ denotes L_2-norm.

It is worth noting that when computing first-order and second-order attention within the Transformation Layers, a multi-head attention strategy can be employed to learn distinct pedestrian features from different subspaces. To allow different heads to learn diverse pedestrian features, we impose an orthogonal loss on the multi-head attention feature map of the last Transformer Layer A' as follows:

$$\mathcal{L}_{omha} = \|\hat{A}'\hat{A}'^{T} - I_m\|_F, \tag{19}$$

where $\hat{A}'$ is obtained by performing L_2 norm to A'; m is the number of heads; I_m is an $m \times m$ identity matrix; $\|\cdot\|_F$ represents the Frobenius norm. Similarly, for the same purpose, the orthogonal loss is also applied to the class tokens of the last Transformer Layer:

$$\mathcal{L}_{omct} = \|\hat{y}_{cls}\hat{y}_{cls}^{T} - I_M\|_F, \tag{20}$$

where $\hat{y}_{cls}$ is derived by performing L_2 norm row-wise to y_{cls}. Furthermore, we impose a cross-entropy loss on the multilayer aggregated features. Similar to DPM [55], we introduce an extra angular margin in softmax function [57] when calculating the predicted labels:

$$\mathcal{L}_{agg} = -\sum_{i=1}^{U} p_i \log \frac{e^{\varepsilon \cdot cos(\theta_i + b)}}{e^{\varepsilon \cdot cos(\theta_i + b)} + \sum_{j=1, j\neq i}^{U} e^{\varepsilon \cdot cos\theta_j}}, \tag{21}$$

where $\theta_i =< F_{egf}, \varphi(F_{agg})^i >$, φ denotes the linear transform; b denotes the angular margin and ε is a scaling factor.

Eventually, the overall objective function can be formulated as follows:

$$\mathcal{L}_{total} = \frac{\alpha}{M+1}\sum_{i}^{M+1}\mathcal{L}_{ce}^{i} + (1-\alpha)\mathcal{L}_{agg} + \frac{1}{M+1}\sum_{i}^{M+1}\mathcal{L}_{tri}^{i} + \beta\mathcal{L}_{omct} + \frac{(1-\beta)}{M}\sum_{j}^{M}\mathcal{L}_{omha}^{j}, \tag{22}$$

where α and β serve as hyper-parameters to balance different classification losses and orthogonal losses, respectively.

IV. Experiments

In this section, we start by introducing the experimental setups, including the benchmark dataset, evaluation protocols, and implementation details. Then, a comparative comparison is conducted between the proposed method and state-of-the-art methods on both occluded and holistic person ReID datasets. Finally, we conduct comprehensive ablation studies and algorithm analysis to demonstrate the effectiveness of the essential components in our proposed method.

A. Experimental Setups

1) Datasets: We validate our method on occluded and holistic person re-identification datasets. *Occluded-DukeMTMC* [20] is derived from *DukeMTMC-reID* [23], encompassing 15,618 training images with 201 identities, 2,210 query images with 519 identities, and 17,661 gallery images with 1,110 identities, where all the query images are occluded. *Occluded-REID* [21] contains 100 occluded persons and 100 holistic persons with 200 identities. Two protocols are employed for this dataset: one using Market-1501 as the training set and Occluded-REID as the test set [7], [42], [47], [48], and the other utilizing half of Occluded-REID for training and the remaining half for testing [21], [26], [43], [44].

Additionally, *Market-1501* [22] comprises 12,936 training images with 751 identities, 3,368 query images, and 15,913 gallery images with 750 identities. *DukeMTMC-reID* [23] consists of 16,522 training images with 702 identities, 2,228 query images, and 17,661 testing images with 702 identities. *CUHK03-NP* [24] is divided based on a new testing protocol [58], including 13,164 images with 1,467 identities,

of which 767 identities are used for training and the remaining 700 identities for testing. These images are categorized into manually labeled bounding boxes (Labeled) and automatically detected bounding boxes using a deformable part model detector [59] (Detected).

2) Evaluation Protocols: Following the standard experimental protocol for person ReID, we employ the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as evaluation metrics. The CMC assesses the recognition accuracy of all query images at various rankings, while the mAP reflects the average of the mean precision of all query images, which can comprehensively measure the model performance.

3) Implementation Details: We adopt the ViT [16] model pre-trained on the ImageNet as the backbone architecture comprising 12 transformer layers. We set the number of patches N_0 to 242, the dimension R to 768, the number of class tokens M to 5, and the hyper-parameter λ to 3. All input images are resized to 256×128 and random horizontal flipping, random cropping, random erasing [60], random grayscale [61], and padding are applied as data augmentation for training. The batch size is set to 64, with 16 identities randomly selected per batch and each identity having 4 images. We use Stochastic Gradient Descent (SGD) as the optimizer with a momentum of 0.9, weight decay of 1e-4, an initial learning rate of 0.008, and apply a cosine learning rate decay strategy. In the objective function, the hyper-parameter ε and b are set to 30 and 0.5, respectively. The hyper-parameter α and β are set to 0.5 and 0.9, respectively. All experiments are conducted on a server with an NVIDIA RTX 4090 GPU.

B. Experimental Results

1) Results on Occluded Datasets: We compare the proposed method with existing ReID methods on two occluded datasets: Occluded-DukeMTMC [20] and Occluded-REID [21]. As shown in Table I and Table II, the comparative methods consist of 10 methods specially designed for CNN-based person ReID, independent of supplementary cues: PCB [8], AFPB [21], FPR [42], Teacher-S [43], IGOAS [44], QPM [45], NetVLAD-M [26], PRE-Net [10], MHSA-Net [62] and RTGAT [46],12 methods introducing additional cues onto CNN or Transformer networks, primarily covering person pose information and semantic parsing: PGFA [20], PEFB [12], RFCNet [63], HOReID [47], PVPM+Aug [7], LKWS [48], FRT [15], BPBreID [49], ISP [11], HCGA [14], MSDPA [50], and Pirt [51], and 6 Transformer-based approaches: AAformer [40], TransReID [17], PAT [52], DRL-Net [53], AMG [54], and SCAT [64].

For the Occluded-DukeMTMC dataset, as shown in Table I, our proposed OAT achieves the best results in terms of mAP (62.2%), Rank-1 (71.8%), and Rank-10 (87.1%), and it achieves the second-best results on Rank-5 (83.0%), merely following behind that of HCGA (83.3%). OAT improves mAP by 0.5% and Rank-1 by 1.4% compared with the second-best method MSDPA [50]. Notably, HCGA and MSDPA both introduce additional cues for network training, while OAT relies solely on the Transformer architecture and introduces two novel modules, yielding superior performance. It can also be observed that the performance of other Transformer-based methods falls significantly behind our method, demonstrating the effectiveness of the proposed SOA and EGF modules.

For the Occluded-REID dataset, as shown in Table II, when adopting Market-1501 as the training set and Occluded-REID as the test set, OAT achieves the best results in terms of mAP (78.2%), Rank-1 (82.6%), outperforming the second-best method MSDPA [50] by 0.7% in both mAP and Rank-1. Furthermore, when employing half of Occluded-REID for

TABLE I

QUANTITATIVE COMPARISON OF OUR OAT WITH STATE-OF-THE-ART METHODS ON OCCLUDED-DUKEMTMC (%). THE COMPETITORS ARE CATEGORIZED INTO THREE TYPES: CNN-BASED, EXTERNAL CUES-BASED, AND TRANSFORMER-BASED METHODS. **BOLD** INDICATES THE BEST RESULTS

Method	Reference	mAP	Rank-1	Rank-5	Rank-10
PCB [8]	ECCV18	33.7	42.6	57.1	62.9
IGOAS [44]	TIP21	49.4	60.1	-	-
QPM [45]	TMM22	49.7	64.4	-	-
PRE-Net [10]	TCSVT23	55.2	68.3	82.7	86.2
MHSA-Net [62]	TNNLS23	44.8	59.7	74.3	79.5
RTGAT [46]	TIP23	50.1	61.0	69.7	73.6
PGFA [20]	ICCV19	37.3	51.4	68.6	74.9
HOReID [47]	CVPR20	43.8	55.1	-	-
ISP [11]	ECCV20	52.3	62.8	78.1	82.9
Pirt [51]	ACM MM21	50.9	60.0	-	-
PEFB [12]	TNNLS22	43.5	56.3	72.4	78.0
RFCNet [63]	TPAMI22	54.5	63.9	77.6	82.1
MSDPA [50]	ACM MM22	61.7	70.4	-	-
FRT [15]	TIP22	61.3	70.7	-	-
BPBreID [49]	WACV23	54.1	66.7	-	-
HCGA [14]	TIP23	57.5	70.2	**83.3**	87.0
AAformer [40]	ArXiv21	58.2	67.0	81.5	-
TransReID [17]	ICCV21	59.2	66.4	-	-
PAT [52]	CVPR21	53.6	64.5	-	-
DRL-Net [53]	TMM23	50.8	65.0	79.3	83.6
AMG [54]	TMM23	59.7	68.5	-	-
SCAT [64]	TII24	54.9	62.8	78.1	83.1
OAT (Ours)	-	**62.2**	**71.8**	83.0	**87.1**

TABLE II

QUANTITATIVE COMPARISON OF OUR OAT WITH STATE-OF-THE-ART METHODS ON OCCLUDED-REID (%). THE COMPARISON REFERS TO TWO PROTOCOLS. THE FIRST GROUP: METHODS THAT EMPLOY MARKET-1501 AS THE TRAINING DATASET AND OCCLUDED-REID AS THE TEST SET. THE SECOND GROUP: METHODS THAT USE HALF OF THE OCCLUDED-REID FOR TRAINING AND THE REMAINING HALF FOR TESTING. **BOLD** INDICATES THE BEST RESULTS

Method	Reference	mAP	Rank-1	Rank-5
PCB [8]	ECCV18	38.9	41.3	-
FPR [42]	ICCV19	68.0	78.3	-
HOReID [47]	CVPR20	70.2	80.3	-
PVPM+Aug [7]	CVPR20	61.2	70.4	84.1
PAT [52]	CVPR21	72.1	81.6	-
LKWS [48]	ICCV21	71.0	81.0	-
MSDPA [50]	ACM MM22	77.5	81.9	-
FRT [15]	TIP22	71.0	80.4	-
RTGAT [46]	TIP23	51.0	71.8	**94.6**
BPBreID [49]	WACV23	68.6	76.9	-
OAT (Ours)	-	**78.2**	**82.6**	91.7
AFPB [21]	ICME18	-	68.1	88.3
Teacher-S [43]	ArXiv19	77.9	73.7	92.9
IGOAS [44]	TIP21	-	81.1	91.6
NetVLAD-M [26]	TIFS22	84.2	89.8	-
HCGA [14]	TIP23	-	88.0	96.0
OAT* (Ours)	-	**89.3**	**94.2**	**97.6**

TABLE III
QUANTITATIVE COMPARISON OF OUR OAT WITH STATE-OF-THE-ART METHODS ON MARKET-1501 AND DUKEMTMC-REID (%). THE FIRST GROUP: HOLISTIC PERSON REID METHODS. THE SECOND GROUP: OCCLUDED PERSON REID METHODS. **BOLD** INDICATES THE BEST RESULTS

Method	Reference	Market-1501		DukeMTMC-reID	
		mAP	Rank-1	mAP	Rank-1
PCB+RPP [8]	ECCV18	81.6	93.8	69.2	83.3
MGN [31]	ACM MM18	86.9	95.7	78.4	88.7
Auto-ReID [27]	ICCV19	85.1	94.5	-	-
Pyramid [32]	CVPR19	88.2	95.7	79.0	89.0
DSA-reID [38]	CVPR19	87.6	95.7	74.3	86.2
PyrAttNet [35]	TIP20	87.6	**95.8**	78.3	88.4
ISP [11]	ECCV20	88.6	95.3	80.0	89.6
RGA-SC [37]	CVPR20	88.1	**95.8**	74.9	86.1
TransReID [17]	ICCV21	88.9	95.2	82.0	90.7
HAT [41]	ACM MM21	89.5	95.6	81.4	90.4
AAformer [40]	ArXiv21	87.7	95.4	80.0	90.1
NetVLAD-M [26]	TIFS22	88.9	95.5	81.3	90.7
PGFA [20]	ICCV19	76.8	91.2	65.5	82.6
HOReID [47]	CVPR20	84.9	94.2	75.6	86.9
PAT [52]	CVPR21	88.0	95.4	78.2	88.8
IGOAS [44]	TIP21	84.1	93.4	75.1	86.9
Pirt [51]	ACM MM21	86.3	94.1	77.6	88.9
PEFB [12]	TNNLS22	81.3	92.7	72.6	86.2
RFCNet [63]	TPAMI22	89.2	95.2	80.7	90.7
MSDPA [50]	ACM MM22	89.5	95.4	**82.8**	90.9
FRT [15]	TIP22	88.1	95.5	81.7	90.5
BPBreID [49]	WACV23	87.0	95.1	78.3	89.6
PRE-Net [10]	TCSVT23	86.0	94.5	76.5	88.9
AMG [54]	TMM23	88.5	95.0	-	-
DRL-Net [53]	TMM23	86.9	94.7	76.6	88.1
HCGA [14]	TIP23	88.4	95.2	-	-
MHSA-Net [62]	TNNLS23	84.0	94.6	73.1	87.3
SCAT [64]	TII24	88.0	95.1	79.8	89.3
OAT (Ours)	-	**89.9**	95.7	82.3	**91.2**

TABLE IV
QUANTITATIVE COMPARISON OF OUR OAT WITH STATE-OF-THE-ART METHODS ON CUHK03-NP (%). **BOLD** INDICATES THE BEST RESULTS

Method	Labeled		Detected	
	mAP	Rank-1	mAP	Rank-1
PCB+RPP [8]	-	-	57.5	63.7
MGN [31]	67.4	66.8	66.0	66.8
Auto-ReID [27]	73.0	77.9	69.3	73.3
Pyramid [32]	76.9	78.9	74.8	78.9
DSA-reID [38]	75.2	78.9	73.1	78.2
PyrAttNet [35]	73.3	76.0	68.5	70.2
ISP [11]	74.1	76.5	71.4	75.2
RGA-SC [37]	76.5	80.4	73.3	77.4
HAT [41]	80.0	82.6	75.5	79.1
AAformer [40]	77.8	79.9	74.8	77.6
NetVLAD-M [26]	76.7	80.4	74.8	79.7
HCGA [14]	75.8	78.3	73.2	76.9
MHSA-Net [62]	72.7	75.6	69.3	72.8
OAT (Ours)	**81.5**	**83.9**	**78.0**	**80.6**

training and the remaining half for testing, OAT consistently outperforms its competitors by a large margin across all evaluation protocols. It achieves the best performance with an mAP of 89.3%, a Rank-1 of 94.2%, and a Rank-5 of 97.6%.

The excellent performance of OAT on occluded datasets demonstrates the effectiveness of the SOA module in establishing higher-order correlations among different body parts or regions in pedestrian images. Additionally, the EGF module effectively mitigates the uncertainty introduced by occlusion by introducing information entropy.

2) Results on Holistic Datasets: To further validate the robustness of our method, we also evaluate it on the holistic datasets, including Market-1501 [22], DukeMTMC-reID [23], and CUHK03-NP [24]. As shown in Table III and Table IV, the comparative methods include 12 holistic person ReID methods: PCB+RPP [8], MGN [31], Auto-ReID [27], Pyramid [32], DSA-reID [38], PyrAttNet [35], ISP [11], RGA-SC [37], TransReID [17], HAT [41], AAformer [40], and NetVLAD-M [26], and 16 state-of-the-art person occluded ReID methods: PGFA [20], HOReID [47], PAT [52], IGOAS [44], Pirt [51], PEFB [12], RFCNet [63], MSDPA [50], FRT [15], BPBreID [49], PRE-Net [10], AMG [54], DRL-Net [53], HCGA [14], MHSA-Net [62], and SCAT [64].

As indicated in Table III, OAT achieves the best mAP (89.9%) and Rank-1 (91.2%) on Market-1501 and DukeMTMC-reID datasets, respectively. While OAT achieves the second-best mAP and Rank-1 on DukeMTMC-reID and Market-1501 datasets, it merely slightly falls behind the best model by a margin of 0.1% and 0.5%, respectively. Notably, for the CUHK03-NP dataset, it is challenging to train a stable and effective model due to the limited training samples. As shown in Table IV, OAT exhibits significant advantages over its competitors. It achieves the best mAP (81.5%) and Rank-1 (83.9%) on the Labeled and the best mAP (78.0%) and Rank-1 (80.6%) on the Detected, surpassing the second-best model by a margin of 1.5%, 1.3%, 2.5%, and 0.9%, respectively.

The success on the holistic datasets demonstrates that OAT can not only alleviate the impact of occlusion but also enhance the learning of discriminative pedestrian features.

C. Ablation Studies

OAT mainly comprises two novel modules, i.e., the Second-Order Attention module (SOA) and the Entropy Guided Fusion module (EGF). To validate the effectiveness of the components, we conduct ablation experiments on Occluded-DukeMTMC. As shown in Table V, "EGF:✗" indicates that features learned from multiple class tokens are directly averaged without incorporating information entropy to guide feature fusion. "SOA:✗" indicates that in the Transformer Layer, only self-attention (first-order attention) is utilized to extract feature correlations. It can be observed that with the addition of these two modules, the performance of the model is improved. When both the EGF and SOA modules are removed, the performance drops about 1.5% in mAP and 2.4% in Rank-1, respectively. Introducing any of them individually results in performance gains and their collaboration gives the best performance of the model.

We also provide attention map visualizations to qualitatively assess the effectiveness of the SOA and EGF modules. From Fig. 3, it can be observed that while both the first-order and second-order attention modules can focus on the unobstructed human body, the latter exhibits larger scope and intensity. Building on first-order attention, SOA can capture higher-order correlations between human parts or regions, resulting in more comprehensive pedestrian features. In addition, as shown in Fig. 4, simply averaging the multiple class tokens tends to

TABLE V

ABLATION ANALYSIS OF THE PROPOSED COMPONENTS ON OCCLUDED-DUKEMTMC (%). **BOLD** INDICATES THE BEST RESULTS

EGF	SOA	mAP	Rank-1	Rank-5	Rank-10
✗	✗	60.7	69.4	82.1	86.2
✓	✗	61.3	70.1	81.9	86.3
✗	✓	61.8	70.3	82.7	86.6
✓	✓	**62.2**	**71.8**	**83.0**	**87.1**

TABLE VI

INFERENCE SPEED ON OCCLUDED-DUKEMTMC AND MARKET-1501

Method	Inference Time (ms)	
	Occluded-DukeMTMC	Market-1501
OAT w/o SOA	3.25	3.18
OAT w/ SOA	3.57	3.46

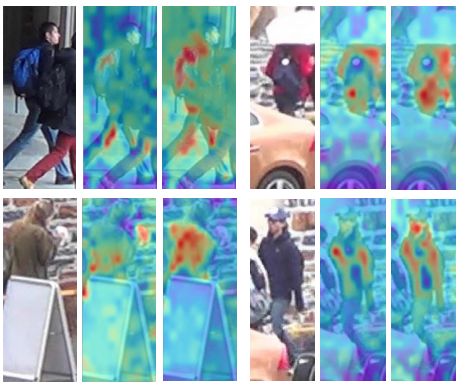


Fig. 3. Qualitative comparison of attention map visualizations between first-order attention and second-order attention. In each group, the 1^{st} image is the original image, the 2^{nd} image shows the results of the first-order attention module, and the 3^{rd} image shows the results of the second-order attention module.

smooth out the attentive features, which inevitably are affected by occlusions. In contrast, utilizing the entropy-guided fusion module can highlight the features of visible human body regions and reduce interference from occlusions. This leads to attention with a larger scope and intensity on the unobstructed human body and produces discriminative pedestrian features.

Additionally, we provide the OAT inference speeds (inference time per image) on the Occluded-DukeMTMC and Market-1501 datasets. As shown in Table VI, "w/o SOA" or "w/ SOA" indicates without or with the SOA module, respectively. We can find that incorporating the SOA module results in only about a 0.3ms increase in inference time, while yielding improvements of 0.9% in mAP (61.3% vs 62.2%) and 1.7% in Rank-1 (70.1% vs 71.8%) on Occluded-DukeMTMC. In summary, the marginal increase in inference time is outweighed by the notable performance enhancements with the inclusion of the SOA module.

D. Algorithm Analysis

1) Influence of the Number of Class Tokens: In Fig. 5, we investigate the impact of varying the number of class tokens, denoted as M, on the performance of our model using Occluded-DukeMTMC and CUHK03-NP. $M = 1$ indicates

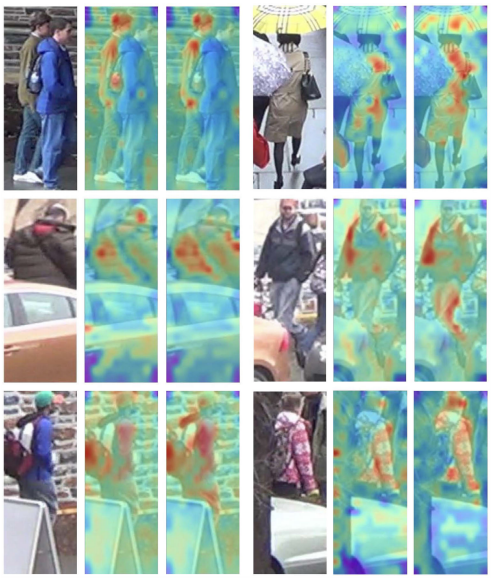


Fig. 4. Qualitative comparison of attention map visualizations with diverse fusion ways for multiple class tokens. In each group, the 1^{st} image is the original image, the 2^{nd} image shows the results of a simple average of multiple class tokens, and the 3^{rd} image shows the results of using entropy guided fusion.

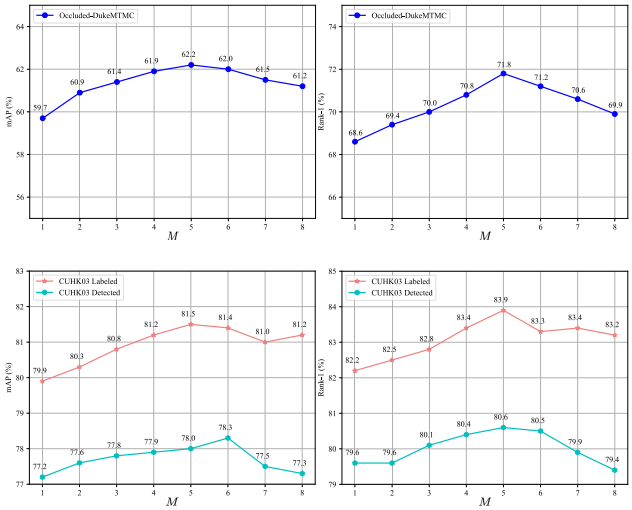
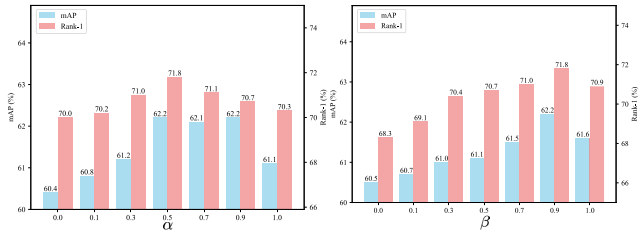


Fig. 5. Ablation analysis of the number of class tokens on Occluded-DukeMTMC and CUHK03-NP datasets.

that only one class token is used to learn global features, which manifests relatively poor performance, obtaining 59.7% mAP and 68.6% Rank-1 respectively on Occluded-DukeMTMC. The performance of the model exhibits a progressive enhancement as the value of M increases. It reaches its peak performance when $M = 5$, achieving 62.2% mAP and 71.8% Rank-1. This suggests that incorporating multiple class tokens enables the learning of various pedestrian features. However, as the value of M further increases, the performance begins to decline. It is possible that adding too many class tokens can lead to feature redundancy, which results in increased difficulty during training, as the model struggles to effectively differentiate between highly similar class tokens. As a result, the model's ability to generalize and discriminate between different classes diminishes, leading to a decline in overall



(a) (b)

Fig. 6. Ablation analysis of hyper-parameters α and β on Occluded–DukeMTMC.

TABLE VII

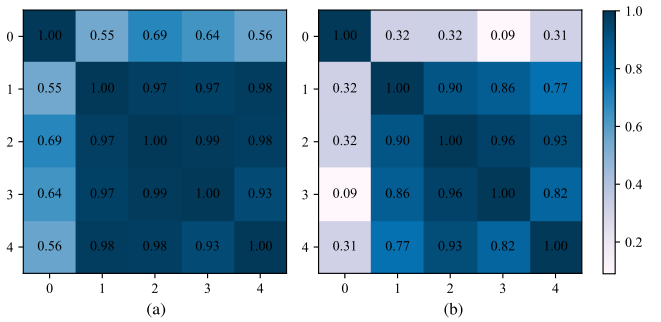
ABLATION ANALYSIS OF THE IMPACT OF EACH LOSS ON OCCLUDED-DUKEMTMC (%). **BOLD** INDICATES THE BEST RESULTS

Method	mAP	Rank-1	Rank-5	Rank-10
OAT w/ce loss	56.7	66.1	81.0	85.3
+triplet loss	59.5	67.0	82.2	86.9
+omct loss	60.4	68.1	82.9	86.7
+omha loss	61.0	70.3	82.9	86.3
+agg loss	**62.2**	**71.8**	**83.0**	**87.1**

performance. A similar trend can be observed on CUHK03, with one exception: the mAP score on CUHK03 Detected is highest when $M = 6$. Thus, we choose $M = 5$ as the default setting.

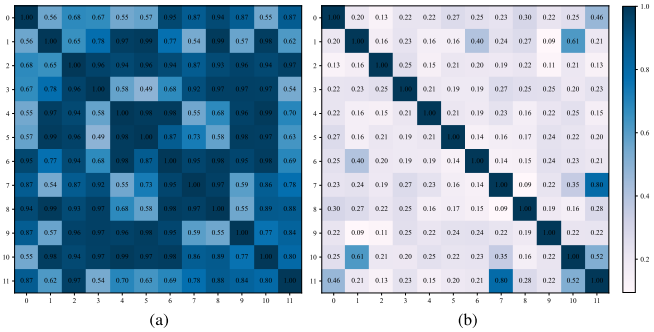
2) Influence of hyper-parameters α and β: The objective function contains two key hyperparameters α and β, which are utilized to balance different classification losses and different orthogonal losses, respectively. We conduct experiments on Occluded-DukeMTMC and show the results in Fig. 6. $\alpha = 0$ indicates that only the multilayer aggregated features are used for classification loss, and $\alpha = 1$ indicates that only the multiple class tokens are used for classification loss. The performance in both of these cases is not satisfactory. As shown in Fig. 6(a), by leveraging the complementary effect of these two classification losses, OAT achieves the optimal performance when $\alpha = 0.5$, resulting in 62.2% mAP and 71.8% Rank-1. We can see a similar trend for parameter β in Fig. 6(b). Imposing the orthogonal loss on both the multiple class tokens and the multi-head attention features leads to an improvement in performance. OAT achieves the peak performance when $\beta = 0.9$.

3) Influence of Different Loss Functions: To thoroughly investigate the impact of each loss function on the model performance, we conduct ablation experiments on Occluded-DukeMTMC. For each loss function, we gradually set its coefficient to 0 to simulate its removal. As shown in Table VII, the model performance gradually improves with the progressive introduction of each loss function. The baseline model merely employs cross-entropy loss for training, yielding 56.7% mAP and 66.1% Rank-1. Upon integrating the triplet loss, the performance improves by 2.8% in mAP and 0.9% in Rank-1. Simultaneous introduction of the orthogonal loss for learning distinct features among multiple class tokens and multi-head attention features improves mAP by 1.5% and Rank-1 by 3.3%. Eventually, incorporating the multi-layer aggregated features for classification loss further boosts



(a) (b)

Fig. 7. Feature correlations between multiple class tokens. (a-b) Training without/with omct loss.



(a) (b)

Fig. 8. Feature correlations between multiple heads. (a-b) Training without/with omha loss.

performance by 1.2% in mAP and 1.5% in Rank-1. This is because the complementary relationship between multilayer aggregated features and multiple class tokens enables the model to capture more discriminative pedestrian features.

To verify the effect of the orthogonal loss more intuitively, we compare the feature correlations of different class tokens and heads without and with orthogonal loss. As illustrated in Fig. 7(a), without using orthogonal loss, there is a high feature correlation between multiple class tokens, which limits the diversity of features. Whereas, as shown in Fig. 7(b), applying orthogonal loss enlarges the dissimilarity of the features, demonstrating that each class token is capable of capturing differentiated pedestrian features. Similarly, applying the orthogonal loss to the multi-head attention features significantly diminishes the inter-head correlations, as shown in Fig. 8. In conclusion, the orthogonal loss effectively enhances diversity between features of multiple class tokens and multiple heads, enabling the model to acquire richer and more discriminative pedestrian features.

V. CONCLUSION

In this paper, we propose a novel person ReID method by incorporating two occlusion-aware modules into Transformer architecture, named OAT. It does not require handcrafted partitions or external cues for training, and can deal with occlusions of diverse types, sizes, etc. The SOA and EGF modules address occlusion challenges by capturing higher-order correlations among different body parts or regions, and by fusing class tokens that are robust to occlusions. Experimental results demonstrate the efficacy of our proposed method in

both holistic and occluded person ReID scenarios. Ablation studies further validate the effectiveness of the component and other design philosophies.

Although OAT achieves an advanced performance compared with other state-of-the-art algorithms, its performance is also susceptible to the coefficient of different loss functions. Besides, the EGF module requires the number of pedestrian categories in a dataset, which brings difficulties in real-world scenes. In the future, we will consider optimizing the network by using adaptive parameter settings.

REFERENCES

[1] P. Fang, J. Zhou, S. K. Roy, P. Ji, L. Petersson, and M. Harandi, "Attention in attention networks for person retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4626–4641, Sep. 2022.

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

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*.

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

[5] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1092–1108, Apr. 2020.

[6] A. Zahra, N. Perwaiz, M. Shahzad, and M. M. Fraz, "Person re-identification: A retrospective on domain specific open challenges and future trends," *Pattern Recognit.*, vol. 142, Oct. 2023, Art. no. 109669.

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

[8] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 480–496.

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

[10] 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.*, vol. 33, no. 8, pp. 4217–4231, Aug. 2023.

[11] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 346–363.

[12] J. Miao, Y. Wu, and Y. Yang, "Identifying visible parts via pose estimation for occluded person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4624–4634, Sep. 2022.

[13] 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 (ICME)*, Jul. 2020, pp. 1–6.

[14] S. Dou, C. Zhao, X. Jiang, S. Zhang, W.-S. Zheng, and W. Zuo, "Human co-parsing guided alignment for occluded person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 458–470, 2023.

[15] 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.

[16] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–11.

[17] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15013–15022.

[18] W. Li et al., "DC-Former: Diverse and compact transformer for person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023, pp. 1415–1423.

[19] X. Ye, W. Zhao, H. Lu, and Z. Cao, "Learning second-order attentive context for efficient correspondence pruning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023, pp. 3250–3258.

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

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

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

[23] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 17–35.

[24] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[25] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–20, Dec. 2017.

[26] M. Zhang et al., "Person re-identification with hierarchical discriminative spatial aggregation," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 516–530, 2022.

[27] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-ReID: Searching for a part-aware ConvNet for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3749–3758.

[28] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019.

[29] R. Quan, Y. Wu, X. Yu, and Y. Yang, "Progressive transfer learning for face anti-spoofing," *IEEE Trans. Image Process.*, vol. 30, pp. 3946–3955, 2021.

[30] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, "Unsupervised person re-identification via cross-camera similarity exploration," *IEEE Trans. Image Process.*, vol. 29, pp. 5481–5490, 2020.

[31] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[32] F. Zheng et al., "Pyramidal person re-identification via multi-loss dynamic training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8514–8522.

[33] J. Xi et al., "Learning comprehensive global features in person re-identification: Ensuring discriminativeness of more local regions," *Pattern Recognit.*, vol. 134, Feb. 2023, Art. no. 109068.

[34] Y. Li, D. Miao, H. Zhang, J. Zhou, and C. Zhao, "Multi-granularity cross transformer network for person re-identification," *Pattern Recognit.*, vol. 150, Feb. 2024, Art. no. 110362.

[35] N. Martinel, G. L. Foresti, C. Micheloni, and N. Martinel, "Deep pyramidal pooling with attention for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 7306–7316, 2020.

[36] G. Chen, T. Gu, J. Lu, J. Bao, and J. Zhou, "Person re-identification via attention pyramid," *IEEE Trans. Image Process.*, vol. 30, pp. 7663–7676, 2021.

[37] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3186–3195.

[38] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 667–676.

[39] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7297–7306.

[40] K. Zhu et al., "AAformer: Auto-aligned transformer for person re-identification," 2021, *arXiv:2104.00921*.

[41] G. Zhang, P. Zhang, J. Qi, and H. Lu, "HAT: Hierarchical aggregation transformers for person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 516–525.

[42] H. Lingxiao, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8449–8458.

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

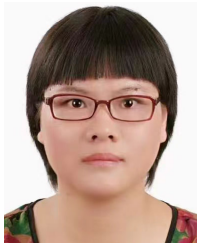
[44] C. Zhao, X. Lv, S. Dou, S. Zhang, J. Wu, and L. Wang, "Incremental generative occlusion adversarial suppression network for person ReID," *IEEE Trans. Image Process.*, vol. 30, pp. 4212–4224, 2021.

[45] P. Wang, C. Ding, Z. Shao, Z. Hong, S. Zhang, and D. Tao, "Quality-aware part models for occluded person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 3154–3165, 2023.

[46] M. Huang, C. Hou, Q. Yang, and Z. Wang, "Reasoning and tuning: Graph attention network for occluded person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 1568–1582, 2023.
[47] G. Wang et al., "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6449–6458.
[48] J. Yang et al., "Learning to know where to see: A visibility-aware approach for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 11885–11894.
[49] V. Somers, C. D. Vleeschouwer, and A. Alahi, "Body part-based representation learning for occluded person re-identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1613–1623.
[50] X. Cheng, M. Jia, Q. Wang, and J. Zhang, "More is better: Multi-source dynamic parsing attention for occluded person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 6840–6849.
[51] Z. Ma, Y. Zhao, and J. Li, "Pose-guided inter-and intra-part relational transformer for occluded person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1487–1496.
[52] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2898–2907.
[53] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 1294–1305, 2023.
[54] J. Mao, Y. Yao, Z. Sun, X. Huang, F. Shen, and H.-T. Shen, "Attention map guided transformer pruning for occluded person re-identification on edge device," *IEEE Trans. Multimedia*, vol. 25, pp. 1592–1599, 2023.
[55] L. Tan, P. Dai, R. Ji, and Y. Wu, "Dynamic prototype mask for occluded person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 531–549.
[56] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[57] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.
[58] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with K-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1318–1327.
[59] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[60] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.
[61] Y. Gong, Z. Zeng, L. Chen, Y. Luo, B. Weng, and F. Ye, "A person re-identification data augmentation method with adversarial defense effect," 2021, *arXiv:2101.08783*.
[62] 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.*, vol. 34, no. 11, pp. 8210–8224, Nov. 2023.
[63] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Feature completion for occluded person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4894–4912, Sep. 2022.
[64] H. Fan, X. Wang, Q. Wang, S. Fu, and Y. Tang, "Skip connection aggregation transformer for occluded person reidentification," *IEEE Trans. Ind. Informat.*, vol. 20, no. 1, pp. 442–451, Jan. 2024.

Yanping Li received the M.S. degree in computer science and technology from Hohai University, Nanjing, China, in 2020. She is currently pursuing the D.Eng. degree with the College of Electronic and Information Engineering, Tongji University. Her research interests include computer vision and person re-identification.

Yizhang Liu received the B.S. degree in electronic and information engineering and the master's degree in computer science and technology from Fujian Agriculture and Forestry University, Fuzhou, China, in 2017 and 2020, respectively. He is currently pursuing the D.Eng. degree with the School of Software Engineering, Tongji University, Shanghai. His current research interests include computer vision and image matching.



Hongyun Zhang received the Ph.D. degree in pattern recognition and intelligence systems from Tongji University, Shanghai, China, in 2005. She is currently an Associate Professor with Tongji University. She is the author or coauthor of nearly 60 journal articles and conference proceedings in principal curves, pattern recognition, machine learning, granular computing, and rough sets. Her current research interests include principal curves, pattern recognition, data mining, image retrieval, and granular computing.

Cairong Zhao (Member, IEEE) received the B.Sc. degree from Jilin University, Changchun, China, in 2003, the M.Sc. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Beijing, China, in 2006, and the Ph.D. degree from Nanjing University of Science and Technology, Nanjing, China, in 2011. He is currently a Professor with Tongji University, Shanghai, China. He is the author of more than 30 scientific articles in pattern recognition, computer vision, and related areas. His research interests include computer vision, pattern recognition, and visual surveillance.

Zhihua Wei received the B.S. and M.S. degrees from Tongji University in 2000 and 2005, respectively, and the integrated Ph.D. degree from Tongji University and Lumière Lyon2 University in 2010. She is currently an Associate Professor with Tongji University. Her research interests include machine learning, image processing, and data mining.

Duoqian Miao is currently a Professor and a Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University. He has published more than 200 articles in IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Pattern Recognition*, *Information Sciences*, and *Knowledge-Based Systems*. His research interests include machine learning, data mining, big data analysis, granular computing, artificial intelligence, and text image processing. His representative awards include the Second Prize of Wu Wenjun AI Science and Technology in 2018, the First Prize of Natural Science of Chongqing in 2010, the First Prize of Technical Invention of Shanghai in 2009, and the First Prize of Ministry of Education Science and Technology Progress Award in 2007. He serves as an Associate Editor for *International Journal of Approximate Reasoning*, *Information Sciences*, and *Chinese Association for Artificial Intelligence*. He serves as the President of the International Rough Set Society (IRSS), the Honorary Chair of the CAAI Granular Computing Knowledge Discovery Technical Committee, the Vice Director for MOE Key Laboratory of Embedded System and Service Computing, and the Vice President of Shanghai Artificial Intelligence Society and Shanghai Computer Society.