scientific reports

Check for updates

OPEN

AIRHF-Net: an adaptive interaction representation hierarchical fusion network for occluded person re-identification

Shuze Geng[1], Qiudong Yu[1]✉, Haowei Wang[2] & Ziyi Song[2]

To tackle the high resource consumption in occluded person re-identification, sparse attention mechanisms based on Vision Transformers (ViTs) have become popular. However, they often suffer from performance degradation with long sequences, omission of crucial information, and token representation convergence. To address these issues, we introduce AIRHF-Net: an Adaptive Interaction Representation Hierarchical Fusion Network, named AIRHF-Net, designed to enhance pedestrian identity recognition in occluded scenarios. Our approach begins with the development of an Adaptive Local-Window Interaction Encoder (AL-WIE), which aims to overcome the inherent subjective limitations of traditional sparse attention mechanisms. This innovative encoder merges window attention, adaptive local attention, and interaction attention, facilitating automatic localization and focusing on visible pedestrian regions within images. It effectively extracts contextual information from window-level features while minimizing the impact of occlusion noise. Additionally, recognizing that ViTs may lose spatial information in deeper structural layers, we implement a Local Hierarchical Encoder (LHE). This component segments the input sequence in the spatial dimension, integrating features from various spatial positions to construct hierarchical local representations that substantially enhance feature discriminability. To further augment the quality and breadth of datasets, we adopt an Occlusion Data Augmentation Strategy (ODAS), which bolsters the model's capacity to extract critical information under occluded conditions. Extensive experiments demonstrate that our method achieves improved performance on the Occluded-DukeMTMC dataset, with a rank-1 accuracy of 69.6% and an mAP of 61.6%.

Keywords Occluded, Re-identification, Adaptive interaction, Hierarchical fusion

Current methods tackling occluded person re-identification often utilize external models such as pose estimation and human parsing to enhance accuracy in scenarios of partial visibility[1–4]. While this approach leverages auxiliary data effectively, it considerably increases computational demands, making it less efficient for processing extensive sequences and adding significant computational overhead. The Hierarchical Aggregation Transformers (HAT) for person re-identification, as presented by Zhang et al.[5], introduce a hierarchical structure to aggregate features at multiple scales, which enhances the model's ability to capture both local and global information. However, HAT may still struggle with significant domain shifts, as it does not explicitly address the challenge of adapting to new domains with different background styles.

In contrast, the Local Correlation Ensemble with GCN based on Attention Features for Cross-Domain Person Re-ID, proposed by Zhang et al.[6], focuses on reducing intra-class differences and improving the reliability of class centers. The method incorporates a pedestrian attention module to emphasize person-specific features and a priority-distance graph convolutional network (PDGCN) to refine class center predictions. While this approach effectively leverages unlabeled target domain data and reduces intra-class variations, it may require more computational resources due to the additional complexity of the PDGCN and the need for robust clustering, which can be challenging in practice.

As the field gravitates towards more streamlined models, sparse attention mechanisms based on Vision Transformers (ViTs) have gained prominence in person re-identification (Re-ID)[7]. ViTs dissect images

[1]School of Information Technology and Engineering, Tianjin University of Technology and Education, Tianjin 300350, China. [2]School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China. ✉email: 475282142@qq.com

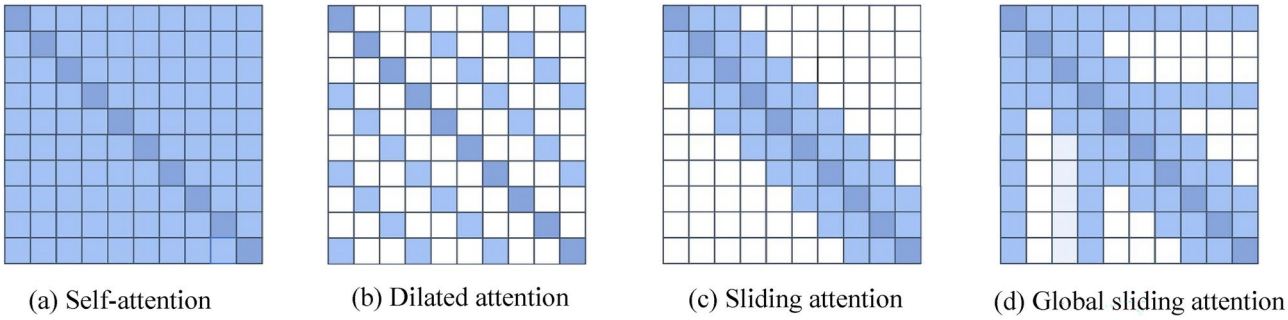

Fig. 1. Self-attention vs. sparse attention patterns (Dark blue squares represent tokens attending to themselves, light blue squares indicate attention computations between the corresponding dark blue square token and other tokens).

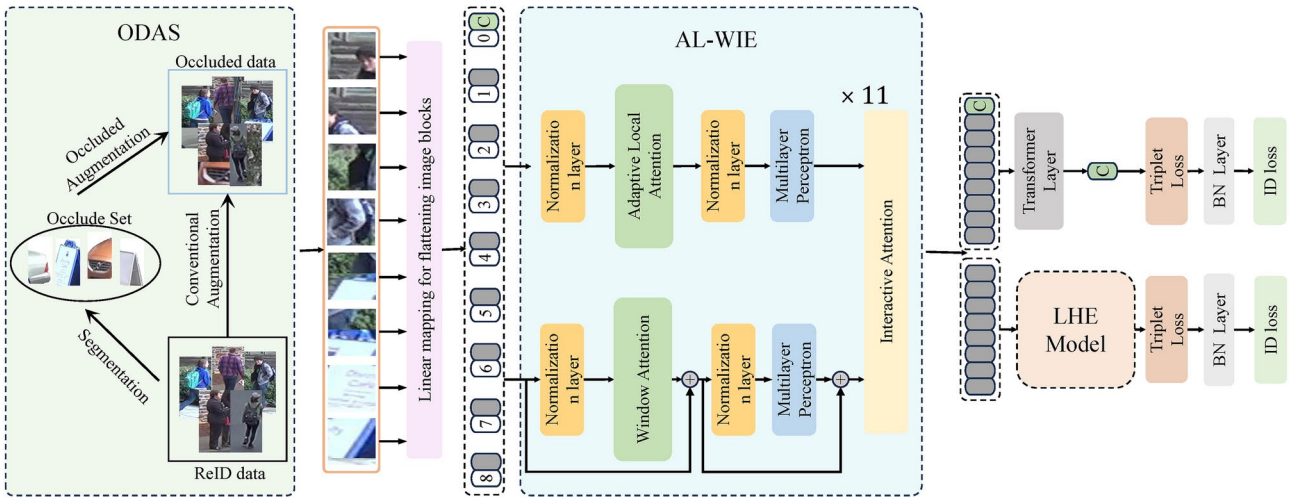

Fig. 2. Overall architecture of the proposed AIRHF-Net. The images of pedestrians in this figure are from the Occluded-DukeMTMC dataset. https://github.com/lightas/Occluded-DukeMTMC-Dataset.

into smaller patches (tokens) and leverage the Transformer architecture to analyze these tokens, capturing comprehensive contextual information. Traditional self-attention mechanisms, which compute interactions among all tokens, lead to a complexity of O(nÂ²) that becomes unsustainable with longer sequences[8,9]. To counter this, sparse attention strategies reduce computational complexity by focusing selectively on crucial tokens, thus significantly lightening the computational load. This selective focus not only streamlines processing but also maintains robustness in recognizing occluded individuals. Figure 1 provides an analysis of the complexity of different sparse attention models. Specifically, (a) represents self-attention without sparsification. Figure 1b shows dilated sparse attention, while (c) and (d) depict different sliding window attentions, where the computational complexity is positively correlated with the number of surrounding tokens being attended to.

Analysis of the figure reveals that manually designed sparse attention methods significantly reduce computational overhead, facilitating model lightweighting. However, practical applications encounter several challenges. Firstly, sparsification may lead to a decline in model performance when processing complex long sequences, and the extraction of key information in specific scenarios may be insufficient, adversely affecting recognition accuracy. Secondly, manually designed sparse attention strategies often suffer from a high degree of subjectivity. Designers must determine which tokens are worth attending to and which can be disregarded, a process that is prone to biases that can result in the omission of important information. Additionally, as model architectures (e.g., ViT) stack layers, the token representations across layers tend to converge, diminishing the model's ability to distinguish features from different spatial locations. The loss of spatial information is particularly detrimental for visual tasks, especially for occluded person re-identification, which requires fine spatial feature differentiation.

To address these issues, this paper proposes a lightweight, adaptive interaction representation hierarchical fusion network, as illustrated in Fig. 2. This network consists of three primary components: an Occlusion-based Data Augmentation Strategy (ODAS), an Adaptive Local-Window Interaction Encoder (AL-WIE), and a Local Hierarchical Encoder (LHE). Firstly, to enhance the extraction of critical information in occluded scenarios while incorporating sparsification, the ODAS is proposed. This strategy increases the number of occluded

samples during model training to improve data diversity and consistency while ensuring the quality of the generated samples. Secondly, compared to the overly subjective manually designed sparse attention strategies, the proposed AL-WIE architecture can adaptively leverage relation-aware mechanisms to extract visible pedestrian regions and contextual features from larger-scale window tokens. This alleviates the issue where, under sparse mechanisms, the model fails to capture all critical contextual information within long sequences, resulting in significant performance degradation. Simultaneously, it filters out occlusion noise and reduces computational complexity. Structurally, the architecture consists of three sub-modules: Large-Scale Window Attention, Association-Aware Adaptive Local Attention, and Interactive Attention Additionally, to combat the issue of progressive spatial information loss resulting from increased stacking of ViT layers, we propose the a Local Hierarchical Encoder (LHE). Unlike methods that segment local features horizontally, the LHE partitions the input sequence along the spatial dimension, integrating features from different spatial areas to generate hierarchical representations of various local features. This strategy effectively reduces the similarity between local features, significantly enhancing their discriminability.

The main contributions of our work can be summarized as follows:

- Firstly, we introduce an AL-WIE that leverages attention maps from ViTs to dynamically extract visible pedestrian regions and contextual features. This is accomplished through three sub-modules: Large-Scale Window Attention, Adaptive Local Attention, and Interaction Attention, which aid in filtering out occlusion noise and reducing computational complexity.
- Secondly, we develop a local hierarchical encoder that partitions the input sequence along the spatial dimension and integrates features from various spatial dimensions to generate hierarchical representations. This reduces the similarity among local features, enhancing their discriminability.
- Thirdly, an occlusion sample augmentation strategy is proposed to enrich person images by substituting different occluders while preserving the identity of the subject, improving both the quality and coverage of the dataset, and facilitating more efficient learning from sparse data.The remainder of this paper is structured as follows: section "Related work" provides a comprehensive review of related works in the field. In section "Methodology", we present the architecture of the proposed framework and elaborate on the implementation details. Experimental results and a thorough analysis of the effectiveness of our method are presented in section "Experimental results and analysis". Finally, section "Conclusion" concludes our work by summarizing the key findings and contributions.

Related work

Due to the relevance of our approach to data augmentation, attention mechanisms, and lightweight methods, this section mainly analyzes algorithms related to these three aspects.

Occlusion augmentation strategies

Existing person re-identification models struggle with occlusions, as the interference caused by them limits robustness. A key factor is the limited number of occluded samples in the training set[2,10], which prevents the model from learning the relationship between occlusions and pedestrians. An effective way to address this issue is through occlusion augmentation strategies.

Zhong et al.[11] introduced a method where pixels are randomly erased and replaced with random values on the image. This approach is simple and helps reduce the risk of overfitting, but it has limited generalization capabilities. As an improvement, Chen et al.[12] randomly cropped rectangular regions from training images, scaled the cropped regions, and randomly pasted them onto one of four predefined areas. Compared to the former strategy, the generated images better simulate real-world occlusion scenarios, enabling the model to implicitly learn more robust features. Similar methods also exist[13,14]. Jia et al.[15] proposed cropping different occluders from the training set and randomly synthesizing them for each training batch. Similar to Chen et al.'s approach, the generated images better simulate real-world occlusion scenarios, allowing the model to implicitly learn more robust features[16]. Likewise, Wang[10]proposed Parallel Augmentation and Dual Enhancement (PADE), which includes a Parallel Augmentation Mechanism (PAM) designed to generate more suitable occlusion data to mitigate the negative impact of imbalanced data. However, using background elements from the dataset for occlusion synthesis differs from real-world scenarios[17]. In contrast to strategies that use only backgrounds as occluders, our method employs actual occluders, making it closer to real-world situations. It also considers the spatial positions of the occluders rather than randomly distributing them, thereby ensuring data complexity and consistency.

Attention and lightweight design

Due to the capability of self-attention mechanisms to model relationships between different semantic component representations on a global scale, some scholars have progressively developed ReID methods based on self-attention[18]. For instance, He et al.[19] proposed a pure transformer framework named TransReID for object re-identification tasks, demonstrating the robustness of self-attention mechanisms. Wang et al.[20] developed a Pose-Guided Feature Disentangling (PFD) framework, which simultaneously trains a pure transformer network and a pose estimation model with learnable parameters. Similar approaches include[21]. Despite their superior performance, the quadratic time complexity of self-attention operations and the incorporation of pose-guided models slow down their execution speed. Recent studies have proposed various methods to tackle this problem, especially in the field of person Re-ID, where these methods have shown considerable potential. For instance, the study on reference[22] introduced a score-based diffusion model to handle incomplete multimodal data in emotion recognition. This approach effectively maps input Gaussian noise to the distribution space of the missing modalities and recovers the data according to their original distributions, while reducing semantic ambiguity

between the recovered data and the missing modalities. Another notable contribution is[23], which proposed an innovative recovery paradigm called DiCMoR to deal with missing modalities in incomplete multimodal learning. DiCMoR maintains the consistency of the recovered data by transferring distributions from available modalities to missing ones, thereby enhancing the model's performance in classification tasks. Similar to this is the study of[24,25]. These studies indicate that by integrating various attention modules and hierarchical network structures, the performance of person re-identification tasks can be effectively improved, especially when dealing with occluded or missing data.

To address the challenges of complexity, recent research has proposed various methods for constructing more efficient transformer architectures[26]. Notably, studies have revealed a degree of sparsity within the self-attention mechanisms of visual transformers, prompting approaches that prune tokens based on their importance scores. Rao et al.[27] introduced DynamicViT (Dynamic Visual Transformer), featuring a lightweight prediction module designed to estimate the importance scores of each token, thereby facilitating the pruning of redundant tokens. Similarly, Meng et al.[28] presented AdaViT (Adaptive Visual Transformer), which strategically learns which blocks, self-attention heads, and token usage strategies should be retained within the transformer's backbone network. Both methodologies incorporate additional lightweight prediction modules that track informative token data during the training phase and consequently exclude this information during inference. Another innovative approach leverages class token attention to retain focused tokens while pruning those that are less informative[29]. Liang et al.[30] proposed EViT (Extended Visual Transformer), defining focused tokens as the image tokens that attract the maximum attention from class tokens. This method merges less informative tokens into a consolidated representation. Building on this, Yin et al.[31] introduced A-ViT (Adaptive Token Visual Transformer), which employs an adaptive token pruning mechanism that dynamically adjusts computational costs based on the complexity of different images. In line with the DyViT approach, the HTS strategy also integrates an additional lightweight prediction module to determine which tokens should be discarded. However, it is crucial to note that the pruned tokens may contain vital features related to human body parts, which can lead to contamination of the feature representation and, consequently, degrade overall performance.

Methodology

In this section, we present the implementation details of our AIRHF-Net method for occluded person re-id, as shown in Fig. 2. The AIRHF-Net, network primarily consists of three components: ODAS, AL-WIE, and LHE.

ODAS

Existing ReID methods struggle to handle occluded images, partly due to a lack of occlusion data. Figure 3a shows that the Occluded-DukeMTMC training set contains a limited number of occluded samples, resulting in insufficient diversity during training, whereas the test set has a higher proportion of occluded samples. Figure 3b illustrates that gallery images are almost unobstructed, while query images are nearly all occluded, leading to inadequate generalization of the model to occlusions. Consequently, researchers[11,32] employed data augmentation strategies to simulate real-world occlusions, such as random cropping and random erasing on individual images. While these techniques enhance the diversity of training data, augmentation performed on individual data points leads to a lack of consistency and makes model training more challenging to converge. Some methods[32] use backgrounds as artificial occlusions. However, there is a notable difference between simple background occlusions and real occlusions. Furthermore, attention modules, which are widely used in re-identification research, naturally focus more on semantically rich foregrounds rather than backgrounds. As a result, the network inevitably overlooks occlusions formed by backgrounds. This oversight leads to the model's inadequate perception of occlusions.

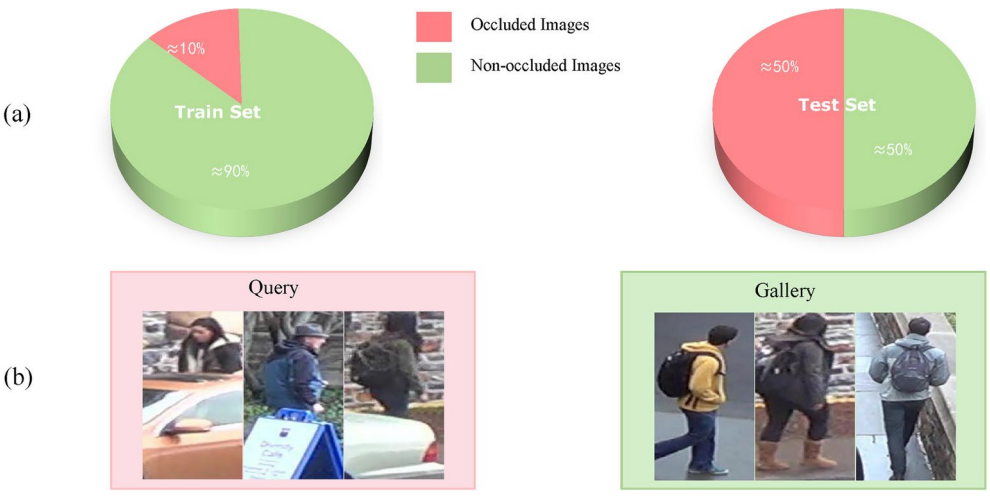

Fig. 3. Current issues in occluded person re-identification datasets.The images of pedestrians in this figure are from the Occluded-DukeMTMC dataset. https://github.com/lightas/Occluded-DukeMTMC-Dataset.

To address the aforementioned issues, this paper adopts an occlusion sample augmentation strategy to enhance pedestrian images. While ensuring the retention of pedestrian identity, the proposed method generates new images by replacing different occluders. Compared to existing similar occlusion augmentation methods, our strategy has two significant differences: First, the occluders (Masks) utilized in our approach are sourced from real occluding objects within the dataset, which have been carefully selected and acquired through manual methods. In contrast, other methods typically utilize random background cropping or randomly erase certain regions of the image to create occlusion Masks. Thus, our approach exhibits a higher level of credibility in simulating real occlusion scenarios. Second, the proposed occlusion sample augmentation strategy considers the spatial positioning of the occluders within the image. Specifically, when designing the occlusion masks, we place the occluders on the left, right, top, and bottom sides of the image. This design approach differs from other methods that randomly generate occluder positions, effectively enhancing data consistency while ensuring occlusion complexity.

Specifically, we first manually segment the occluders covering the pedestrian areas in the training set to create an occlusion $Set_{occluded}$. This effectively ensures the diversity of the occluders and guarantees that they do not appear in the test set. Utilizing $P \times K$ sampling, where *P* pedestrian IDs are randomly selected from the original images and *K* images are randomly chosen for each pedestrian ID. Unlike single data augmentation methods, our strategy selects *K* occluders as augmentation samples for each batch. The same *K* occluders are used to augment images of different identities within the same batch, rather than selecting occlusion samples. This approach helps maintain a certain level of complexity while ensuring data consistency.

Given an input image, we first perform conventional data augmentation, such as resizing, random flipping, padding, and cropping, to obtain the conventionally augmented image $x \in \mathbb{R}^{H \times W \times C}$, which has a size of $S = H \times W$. Next, to simulate realistic occlusion scenarios, the occlusion sample P_o is resized to $S_o = t_o \times S$, where $t_o \sim \mu(0.2, 0.5)$ (with μ representing the uniform distribution), resulting in P_o having a height $H_o = \sqrt{S_o \times t_s}$ and width $W_o = \sqrt{\frac{S_o}{t_s}}$, with $t_s \sim \mu(0.3, 3.3)$ denoting the specific values drawn from this distribution.

To increase the diversity of the occlusion samples, we also apply conventional augmentation techniques to them. Based on empirical observations, occluders tend to appear in four spatial positions relative to the object of interest: top, bottom, left, and right. Therefore, we construct a coordinate set Set_{corner} as follows:

$$\mathrm{Set_{corner}} = \begin{cases} (x_\mathrm{t}, y_\mathrm{t}) & \in & (H - H_\mathrm{o}, H), (0, W), \\ (x_\mathrm{b}, y_\mathrm{b}) & \in & (0, H - H_\mathrm{o}), (0, W), \\ (x_\mathrm{l}, y_\mathrm{l}) & \in & (0, W - W_\mathrm{o}), (0, H), \\ (x_\mathrm{r}, y_\mathrm{r}) & \in & (W - W_\mathrm{o}, W), (0, H), \end{cases} \tag{1}$$

After selecting occlusion sample , we calculate its height-to-width ratio, denoted as $\alpha = H_o/W_o$. If α is less than 2, the sample is considered a horizontal occluder; otherwise, it is deemed a vertical occluder. For horizontal occluders, a random point from the coordinate set $(x_t, y_t)(x_b, y_b)$ is chosen as the starting point, and the augmented occlusion sample is pasted onto the original image to create the final input image $x_{occluded}$ for the network. For vertical occluders, the occluder is placed on either the left or right side of the image, forming a new augmented sample. Figure 4 illustrates the effects of the occlusion sample augmentation strategy, where (a) represents the original image, and (b) shows the input occluded image.

AL-WIE

After obtaining sufficient occlusion samples, it is crucial to reduce the model's attention to the occluders. Due to the quadratic growth in computational complexity of self-attention operations with sequence length in ViTs, the models exhibit high computational complexity. To address this, some methods[33,34] use non-overlapping windows and shifting operations to reduce computational load and achieve global modeling, while others[35]

Fig. 4. Display of occluded images generated by ODAS. The images of pedestrians in this figure are from the Occluded-DukeMTMC dataset. https://github.com/lightas/Occluded-DukeMTMC-Dataset.

design windowed context self-attention mechanisms combined with global self-attention to expand the receptive field. However, the former method may hinder inter-window communication since self-attention is confined within windows. The latter method loses the ability to model multi-scale contexts due to the fixed number of image patches and constant embedding size. To address this, we propose an AL-WIE that reduces the computational complexity of self-attention through sparsification at the attention level and models multi-scale features obtained from different attentions, effectively expanding the model's perception of visible pedestrian areas.

The AL-WIE consists of three submodules: the Large-Scale Window Attention Module, the Association-Aware Adaptive Local Attention Module, and the Interaction Attention Module. Firstly, the Large-Scale Window Attention Module conducts global context interactions over large-scale window image patches to capture the spatial relationships among adjacent windows effectively. Secondly, the Association-Aware Adaptive Local Attention Module exploits the inter-token associative information adaptively to extract pertinent local features from critical image patches. This mechanism mitigates the impact of occlusion while concurrently reducing the computational complexity associated with self-attention operations. Lastly, the Interaction Attention Module integrates the outputs from both branches, facilitating the transmission of inter-window contextual information to the salient image patches within each window. By synthesizing information across various scales, this module significantly enhances the model's perceptual range concerning the visibility of pedestrian areas.

Large-scale window attention

This section follows the setting of ViT, dividing the image into $\{P_0^i\}_i^N$ non-overlapping image patches, and applying linear projections to obtain the embedding patches $X_p \in \mathbb{R}^{N\times W\times C_1}$, where C_1 is the channel dimension. Then, a class token X_{cls} is concatenated to accumulate information from other tokens and serve as the final feature representation for classification. Positional encoding is added to each image patch to help the Transformer model capture the relative positional information of elements in the input sequence, which is crucial for understanding context relationships and modeling long-term dependencies in the sequence. Therefore, the input sequence can be represented as $X = [X_{cls}, X_p] \in \mathbb{R}^{(N+1)\times W\times C_1}$

In self-attention, the input sequence X goes through three linear layers to produce query Q, key K,value V matrices, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{(N+1)\times W\times C_1}$ represents the dimensions of these matrices. The Multi-Head Self-Attention (MSA) operation formula is as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{2}$$

Figure 5 illustrates the structure of the AL-WIE. The adaptive local attention branch continues with the output sequence from the previous encoder layer, while in the window attention branch, tokenization operations are used to obtain a larger-scale input sequence. Previous tokenization operations often employed linear mappings, which might not capture low-level features such as edges and corners in images. To better extract low-level features of image patches, the tokenization operation in the window attention introduces variable convolution kernels and is applied at different stages of the network. This design helps to enhance the network's generalization ability, allowing it to better adapt to images of varying scales and complexities. For instance, if the window token

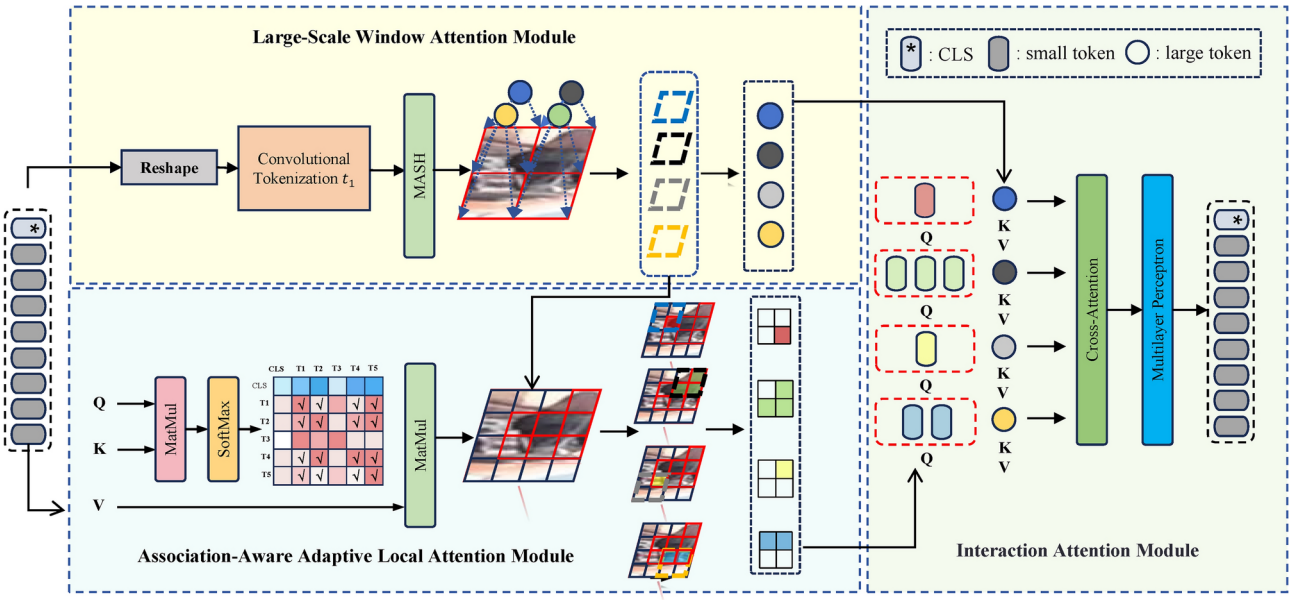

Fig. 5. Architecture of the AL-WIE. The images of pedestrians in this figure are from the Occluded-DukeMTMC dataset. https://github.com/lightas/Occluded-DukeMTMC-Dataset.

size is ${S_{window}}^2$, and the token size in the original input sequence is ${S_{input}}^2$, this means that one window token spatially includes $num_o = (S_{window}/S_{window})^2$ original input tokens locally.

Based on the above analysis, in the window attention branch, the input $\left\{P_0^i\right\}_i^N$ is resized to $C \times H \times W$ through a reshape operation. Subsequently, through the variable convolution tokenization operation t_1, the three-dimensional image is divided into window tokens via convolutional operations.

$$P_W = Conv(Reshape(\left\{P_0^i\right\}_i^N)) \tag{3}$$

The image patches in the window attention branch are denoted as $P_W = \left[p_w^1, p_w^2, ..., p_w^{N/num_0}\right]$, where interactions among all window tokens take place to effectively learn global information while keeping the number of window tokens minimal. This information covers the contextual understanding of the entire image. Since the windows are non-overlapping, this branch can also facilitate information exchange among windows. Consequently, the computation of window features through window attention is given by equation ,

$$F_{w-MSA} = f^w(P_w) + \mathrm{MSA}(LN(f^w(P_w))) \tag{4}$$

where $f^w()$ represents the feature mapping of aligned dimensions, MSA stands for MSA, and LN denotes Layer Normalization.

Association-aware adaptive local attention

In the study of occluded pedestrian re-id, the diversity of backgrounds and occlusions often leads to a significant decline in model performance. To improve performance, the complexity of models has consistently increased, consequently raising the computational costs for training and inference. Inspired by the research of Rao et al.[27], which found that pruning away less informative parts, such as small backgrounds, has little impact on model accuracy, this paper proposes an adaptive sparsification mechanism at the attention level. This approach helps to reduce unnecessary information interference, thereby accelerating training speed and enhancing the effectiveness of occluded pedestrian re-identification.

The module consists of two components: one is the Relation-Aware Adaptive Token Selection module; the other is Sparsification.

Relation-aware adaptive token selection module. In ViTs, all image patches are treated as tokens, and MSA is used to compute the relationships between every pair of tokens. This inevitably results in computational redundancy, as not all image patches are meaningful in the MHSA. Furthermore, Caron et al.[36] demonstrated that the class token in ViT tends to focus more on target regions than on non-target areas in the image. To address this, the present module selects relevant target tokens and filters out irrelevant tokens based on the correlations between image patches and their contributions to recognition. The selected target tokens are fed into subsequent encoders for information interaction, effectively preventing the propagation of irrelevant tokens in the subsequent transformer layers. The proposed module surpasses CNNs by dynamically adjusting its focus according to input features, enhancing flexibility. Unlike fixed convolution kernels, it targets specific relevant regions, capturing essential information more accurately. By accounting for interrelationships between image tokens, it reduces noise and improves performance. This method effectively integrates global information and captures long-range dependencies, enhancing semantic understanding. This approach avoids the introduction of additional noise from frequently selecting occluded image patches and utilizes only the discriminative image patches, which can significantly enhance model performance. The proposed Token Selection module introduces several improvements built upon the ViT framework.

Let $Z \in \mathbb{R}^{(N+1)\times C_1}$ be the input sequence of the ViT, where $N+1$ denotes the sequence length. The first image patch is the class token x_{cls}, representing global features, while the remaining image patches are referred to as image tokens image token $x_{img_i}, i = [1, 2, ..., N]$.

$$\begin{aligned} Z_{cls} &= x_{cls}; \\ Z_{img} &= [x_{img_1}; x_{img_2}; ...; x_{img_N}] \end{aligned} \tag{5}$$

The interaction between the class token and image tokens occurs through the attention mechanism in ViT:

$$\mathrm{Attention}_{cls}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{cls}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} = \boldsymbol{\alpha} \bullet \mathbf{V} \tag{6}$$

Here $\boldsymbol{\alpha} \in \mathbb{R}^{1\times N}$, (excluding the attention values of the class token itself) corresponds to the attention values from the first row to in the attention matrix, as described in Eq. (2). $\mathbf{Q}_{cls}$ denotes the query matrix obtained through the linear mapping of x_{cls} . Since v_i originates from the i-th token of $\mathbb{V}$, the attention value α_i (i.e., the i-th element of $\boldsymbol{\alpha}$) determines how much information from the i-th token is fused into x_{cls}. Therefore, α_i can be interpreted as representing the importance of the i-th token for x_{cls}.

The traditional approaches entail computing the average attention score ($\overline{\alpha}$), across all attention heads. These scores are then ordered in descending order, and a top-k strategy is employed to select the top k image patches

with the highest attention values as target tokens. The remaining image patches are classified as irrelevant tokens, which predominantly include background features and occlusions. This also results in obtaining the indices idx_{obj} and idx_{occ} corresponding to the target and irrelevant tokens. However, in the context of sparse attention mechanisms for processing long sequences, the strategy of focusing solely on the top K important tokens may lead to significant limitations. First, this approach may prevent the model from effectively capturing all relevant contextual information within the long sequence, thereby affecting its performance. Secondly, the selected target tokens may exhibit a high correlation with certain irrelevant tokens, resulting in the model collecting a substantial amount of interference information from these unrelated tokens. Therefore, this paper considers the correlations not only between the class token and image tokens but also between the image tokens themselves.

Specifically, the first column of the attention matrix ω (excluding the attention value of the class token itself) is denoted as $\overline{\beta}$, which represents the attention scores for the interaction of contextual information among image tokens. Thus, the interaction attention values γ are computed as follows:

$$\gamma = \overline{\alpha_i} \bullet \overline{\beta_i}, i = 1, 2, ..., N \tag{7}$$

After obtaining the interaction attention values γ, the top k image patches with the highest attention scores are selected as target tokens, following the top-k strategy. The retention rate of the target tokens is defined as Γ :

$$\Gamma = k/N \tag{8}$$

This module does not incorporate any additional parameters, thereby not increasing the parameter count of the model.

Sparsification

First, the adaptive selection module obtains the indices idx_{obj} and idx_{occ} corresponding to the target tokens and irrelevant tokens. The adaptive local attention module computes the global attention exclusively among the target tokens. After a linear mapping, the target tokens are transformed into query, key, and value matrices, which are calculated using the following formula:

$$\mathbf{Q}_{AL} = P_O^i W_q, \ \ i \in idx_{obj} \tag{9}$$

$$\mathbf{K}_{AL} = P_O^i W_k, \ \ i \in idx_{obj} \tag{10}$$

$$\mathbf{V}_{AL} = P_O^i W_v, \ \ i \in idx_{obj} \tag{11}$$

Where W_q,W_k,W_v, denote learnable linear mapping functions.

$$\mathbf{Q}_{AL} = \left[q_{AL}^1, q_{AL}^2, \ldots, q_{AL}^{\Gamma \times N}\right] \tag{12}$$

$$\mathbf{K}_{AL} = \left[k_{AL}^1, k_{AL}^2, \ldots, k_{AL}^{\Gamma \times N}\right] \tag{13}$$

$$\mathbf{V}_{AL} = \left[v_{AL}^1, v_{AL}^2, \ldots, v_{AL}^{\Gamma \times N}\right] \tag{14}$$

Compared to traditional self-attention operations, the adaptive local attention module significantly reduces the amount of noise information by performing attention calculations exclusively among the target tokens, informed by an interpretive analysis of the attention maps of ViT when handling occlusions[37]. Furthermore, to avoid introducing excessive interfering feature information into the target tokens, this module cleverly omits the residual connection structure. As a result, the output features contain far less noise information than the input features, enabling the model to capture key features more accurately in occluded person re-identification scenarios. Therefore, the computation of adaptive local attention is as follows:

$$\mathrm{F}_{AL-MSA} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{AL}\mathbf{K}_{AL}^T}{\sqrt{d}}\right)\mathbf{V}_{AL} \tag{15}$$

Interactive attention

In order to incorporate global information encompassing the relationships among larger-scale image patches into the highly responsive local areas of smaller-scale patches, thereby enhancing spatial discriminative cues and expanding the model's perceptual range, this section introduces the concept of Interactive Attention. Specifically, following the acquisition of outputs from both Adaptive Self-Attention and Window Attention, to facilitate communication between window tokens and their original counterparts, the globally refined features extracted via Window Attention serve as a bridge to be relayed back to the original tokens within the window. This process aims at mitigating the limitations imposed by single-scale features and further diversifying the representation of pedestrians.

Each window token p_w^i corresponds to num_o original tokens $P_{w-o} = \left[p_o^{i,1}, p_o^{i,2}, ..., p_o^{i,num_o}\right]$ in the spatial dimension, where *i*-th original token within the *j*-th window. Instead of interacting with the tokens within the window, the window token p_w^i performs self-attention operations solely with the target tokens $p_o^{i,j}, j \in idx_{obj}$, following the adaptive local attention framework. This design allows global contextual information to be transmitted only to the tokens containing pedestrian information, thereby reducing the

similarity between the tokens containing pedestrian information and those containing noise information, such as occlusions, which enhances the network's focus on the target pedestrian areas.

Specifically, the query vector $\mathbf{Q}_o^{i,j} = \left[q_o^{i,1}, q_o^{i,2}, ..., q_o^{i,j} j \in idx_{obj}\right]$ is computed from the original tokens $p_o^{i,j}, j \in idx_{obj}$ within the window. The key vector $\mathbf{K}_w^i = \left[k_w^1, k_w^2, ..., k_w^{N/num_o}\right]$ and value vector $\mathbf{V}_w^i = \left[v_w^1, v_w^2, ..., v_w^{N/num_o}\right]$ are computed from the window tokens, as illustrated in the following process:

$$\mathbf{Q}_o^{i,j} = \mathrm{BN}\left[\mathrm{linear}(F_{\text{AL-MSA}})\right] \tag{16}$$

$$\mathbf{K}_w^i = \mathrm{BN}\left[\mathrm{linear}(F_{\text{AL-MSA}})\right] \tag{17}$$

$$\mathbf{V}_w^i = \mathrm{BN}\left[\mathrm{linear}(F_{\text{AL-MSA}})\right] \tag{18}$$

where BN denotes a batch normalization layer, and linear represents a linear layer. The vectors $\mathbf{Q}_o^{i,j}, \mathbf{K}_w^i$, undergo matrix multiplication to obtain the adaptive local-window weights, which are then normalized using a Softmax layer. The normalized weights are multiplied by the value vector $\mathbf{V}_w^i$ to yield the adaptive-window attention interaction features F_{AL-w}^i within a single window, as described in Eq. (19):

$$\mathrm{F}_{AL-w}^i = \mathrm{Softmax}\left(\frac{\mathbf{Q}_o^{i,j}\mathbf{K_w^i}^{\mathbf{T}}}{\sqrt{d}}\right)\mathbf{V}_w^i \tag{19}$$

where *d* represents the dimensionality of the query vector, and *i*, *j* denotes the j-th image patch within the i-th window. Finally, all the obtained adaptive-window attention features are concatenated in the spatial dimension, and a multilayer perceptron is applied to enhance the features, resulting in the final output features F_{AL-w} of the AL-WIE module. It is noteworthy that, aside from layer normalization, the weights between window attention and adaptive local attention are shared. Therefore, the number of parameters does not significantly increase compared to a standard Transformer encoder.

Analysis of complexity

In this section, we discuss the complexity of the AL-WIE, which consists of three components: window attention, Association-Aware adaptive local attention, and interaction attention. We will analyze the complexity of these three components one by one.

For window attention, we divide the original tokens into n segments. Therefore, the window attention includes $\frac{N}{n}$ original tokens, and with n windows, the computational complexity of window attention is given by:

$$O\left(\left(\frac{N}{n}\right)^2 \times n\right)$$

In the case of adaptive local attention, after obtaining the interaction attention values, we select the top k image patches with the highest attention values (top-k strategy) as the target tokens and the retention rate of the target tokens be denoted as Γ. Thus, the computational complexity of adaptive local attention is:

$$O\left((\Gamma \cdot N)^2\right)$$

Interaction attention refers to the interaction between adaptive self-attention and window attention. It is noteworthy that, in this paper, the window attention and adaptive local attention share weights, except for layer normalization. Consequently, the overall computational complexity of interaction attention is:

$$O(\Gamma \cdot n \cdot N)$$

Therefore, the overall time complexity is :

$$O((N^2/n) + (\Gamma N)^2) + \Gamma n N).$$

Furthermore, in this paper, the image size is 256×128, and the patch size is 16×16, The window attention is divided into n=4 segments. This results in N=128, and the retention rate is set to 0.7. Substituting these values into the formulas, we get the following computational complexities: $O(0.74N^2 + 2.8N)$.

Compared to the full self-attention operation, which has a computational complexity of $O(N^2)$, the proposed method reduces the computational complexity by approximately 25%. This reduction in computational complexity demonstrates the efficiency of the AL-WIE, making it a more computationally feasible solution while maintaining a high level of performance.

LHE

Although the ViT excels at utilizing global features to achieve effective recognition performance, in the context of occluded pedestrian re-identification, critical information often relies more heavily on local features. Previous

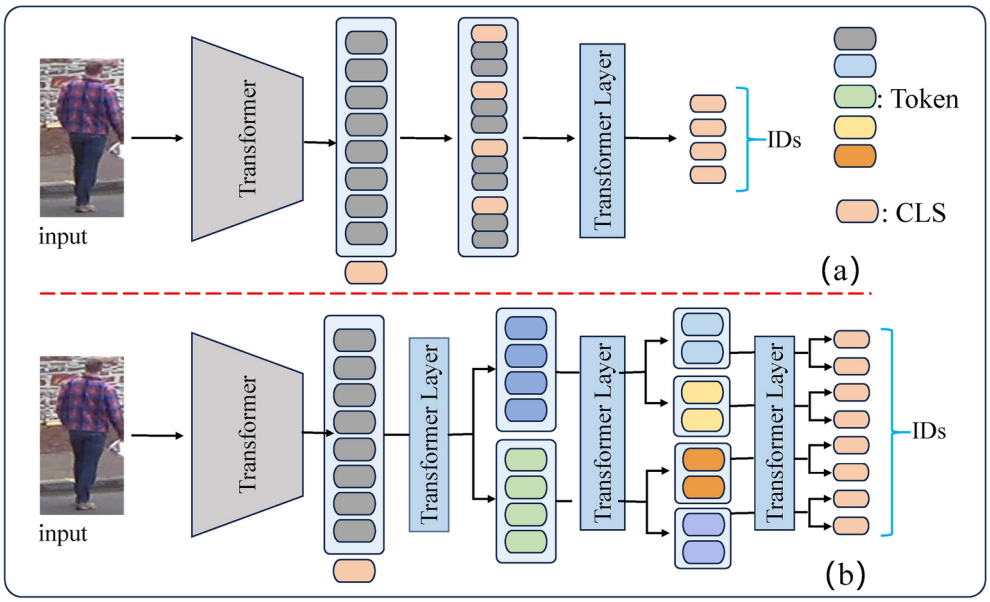

Fig. 6. Improvement diagram of local feature learning module for person re-identification. The images of pedestrians in this figure are from the Occluded-DukeMTMC dataset. https://github.com/lightas/Occluded-DukeMTMC-Dataset.

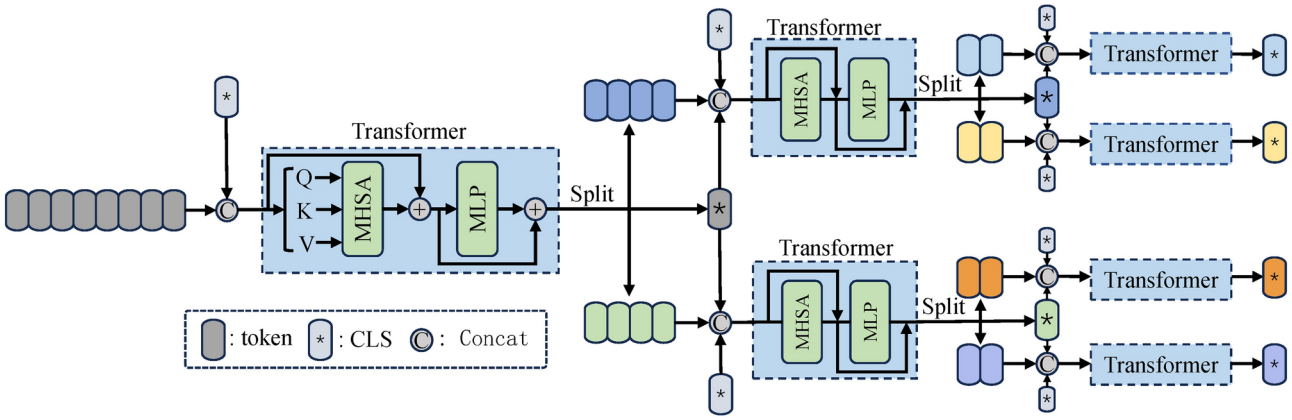

Fig. 7. Architecture of the LHE.

works[16,38] have only horizontally partitioned these local features, as depicted in Figure 6a. However, they fall short in capturing ideal local features. While the input to a Transformer encoder is a one-dimensional sequence, from an image processing perspective, it can be viewed as a sequence of two-dimensional image patches, thus maintaining spatial associations among the sequence elements. In other words, not only do adjacent image patches exhibit strong correlations, but due to the presence of a two-dimensional space, certain patches that are further apart also share a degree of spatial correlation.

Inspired by[39], this paper proposes a LHE, illustrated in Fig. 7. Contrary to the strategy of merely horizontally segmenting local features, the LHE recursively divides the image sequence, extracting spatially correlated features from distinct local regions under the guidance of global semantics. As the hierarchical structure deepens, the diversity of fine-grained cues is captured, significantly enhancing the discriminability of local features.

Specifically, the input to the LHE is denoted as $F_{in} = [X_{cls}, X_1, X_2, ..., X_N]$. First, the class token is separated, resulting in a new sequence of image patches $F_{patch} = [X_1, X_2, ..., X_N]$. Let us assume that the LHE $[LHE^1, LHE^2, ..., LHE^k]$, divides the sequence in LHE_k into 2^k parts, thereby introducing a hierarchical partition for different local features. Specifically, the hierarchical partitioning of the sequence is illustrated as follows:

$$F^1_{patch} = \left[X_{cls}, X_1, X_2, \ldots, X_{N/2^k}\right] \tag{20}$$

$$F_{patch}^{2} = \left[X_{(N+1)/2^k}, X_{(N+2)/2^k}, X_{(N+3)/2^k}, \ldots, X_{(2N)/2^k}\right] \tag{21}$$

$$F_{patch}^{k-1} = \left[X_{(((k-2)N+1)/2^k)}, X_{(((k-2)N+2)/2^k)}, X_{(((k-2)N+3)/2^k)}, \ldots, X_{((k-1)N)/2^k}\right] \tag{22}$$

$$F_{patch}^{k} = \left[X_{(((k-1)N+1)/2^k)}, X_{(((k-1)N+2)/2^k)}, X_{(((k-1)N+3)/2^k)}, \ldots, X_{((k-1)N)/2^k}\right] \tag{23}$$

For the locally segmented image patch sequence $F_{patch}^{i}, i = 1, 2, ..., k$ obtained from the hierarchical partitioning of LHK^k, the class token $X_{cls}^{k-1} \in \mathbb{R}^{1\times C_1}$ output from layer LHK^{k-1} is concatenated with F_{patch}^{i} to guide fine-grained feature learning. Additionally, a new class token X_{cls}^{k} is appended to F_{patch}^{i} to summarize contextual information. Consequently, the input sequence to the k-th local hierarchical encoder is represented as:

$$F_{in}^{i} = \left[X_{cls}^{k-1}, X_{cls}^{k}, F_{patch}^{i}\right], i = 1, 2, ..., k. \tag{24}$$

Subsequently, F_{in}^{i} is processed through a MSA mechanism and a multi-layer perceptron to construct local hierarchical features, with the computational process detailed as follows:

$$\begin{aligned} F_{LHE}^{i\prime} &= f^{LHE}(F_{in}^{i}) + MSA(LN(f^{LHE}(F_{in}^{i}))) \\ F_{LHE}^{i} &= F_{LHE}^{i\prime} + MLP(LN(F_{LHE}^{i\prime})) \end{aligned} \tag{25}$$

Here, $F_{LHE}(\bullet)$ represents a linear mapping, and F_{LHE}^{i} is the output of the local hierarchical encoder . $i = 1, 2, ..., k$.. Then, the class token X_{cls}^{k} from the k outputs of F_{LHE}^{i}. That is separated out as the final local feature representation.

Objective function

This section introduces the loss functions used for training the network. To ensure that the model learns identity information while emphasizing the relative relationships between images, a combination of cross-entropy loss and triplet loss is employed, thereby constructing a multi-task learning framework. Specifically, cross-entropy loss is used to measure the discrepancy between predictions and true labels. In this study, different classifiers are employed to train the local and global features, with each classifier comprising a fully connected layer followed by a Softmax layer, which converts the model outputs into a probability distribution. The formula for cross-entropy loss is as follows:

$$\mathcal{L}_{CE} = -\frac{1}{kB+1}\sum\left(\sum y_i \log\left(\frac{\exp(W_i F_m)}{\sum \exp(W_j F_m)}\right) + y_i \log\left(\frac{\exp(W_i F_G)}{\sum \exp(W_j F_G)}\right)\right) \tag{26}$$

Here, k denotes the number of partitions for local features, B represents the number of images in a batch, W is the linear mapping matrix, and y_i denotes the corresponding labels. F_m and F_G represent the local and global feature representations, respectively. Furthermore, the proposed method utilizes triplet loss to encourage the model to learn discriminative feature representations and to enhance its sensitivity to similarities. The formula for triplet loss is

$$\mathcal{L}_{tri} = \max(0, \mu + \|F_a - F_p\|_2 - \|F_a - F_n\|_2) \tag{27}$$

where F_a, F_p and F_n , refer to the anchor, positive sample, and negative sample, respectively, and μ is a hyperparameter that regulates the distance between positive and negative samples.

In summary, the overall loss function is computed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{tri} \tag{28}$$

Experimental results and analysis

Experimental setup

The network proposed in this paper is deployed in the PyTorch 1.11.0 framework and trained and tested using a single RTX A100 GPU. The method employs ViT as the backbone network, which segments the input images 256×128 into patches of size 16×16. The batch size is set to 32, with each batch containing 4 images of the same identity. Following the standard settings of ViT-base, the backbone network has a depth of $L = 1_2$ layers. In layers 1 to 9, the tokenization operation t_1 segments the images into patches of size 64×64, while in layers 10 and 11, the patch size is set to 32×32 . Additionally, the target token retention rate γ is configured to 0.7 to select image patches with high attention response. Feature learning is constrained using ID loss and triplet loss. For triplet loss, a hard positive and a hard negative sample are selected for each sample in the mini-batch obtained through PK sampling, forming triplets with μ set to 0.3. The number of LHE stacked layers is set to 4. The training process is end-to-end, utilizing the Adam optimizer to minimize the system loss function. The network is trained for 160 epochs with an initial learning rate of 0.008. A warm-up strategy is employed, linearly increasing the learning rate before epoch 5, followed by cosine decay. The model converges within 120

Method	Rank1	Rank5	Rank10	mAP
Baseline	67.1	81.2	85.6	59.4
Random pasting	68.5	82.1	86.9	60.9
Horizontal occlusion	69.2	82.4	87.6	61.3
Vertical occlusion	68.9	82.1	87.4	61.1
Ours	69.6	82.7	88.1	61.6

Table 1. Comparison of different occlusion strategies.

Index	Baseline	ODAS	AL-WIE	LHE	R-1	R-5	R-10	mAP
1	✓				60.7	77.6	82.9	52.3
2	✓	✓			62.4	77.1	83.1	54.8
3	✓		✓		63.5	78.8	83.5	56.3
4	✓			✓	62.8	78.3	82.8	56.7
5	✓	✓	✓		68.4	81.5	85.6	60.5
6	✓	✓		✓	65.6	79.0	83.9	57.5
7	✓		✓	✓	67.1	81.2	85.6	59.4
8	✓	✓	✓	✓	**69.6**	**82.7**	**88.1**	**61.6**

Table 2. Ablation experiments of each module in AIRHF on Occluded-DukeMTMC (%). The best performance values are in bold.

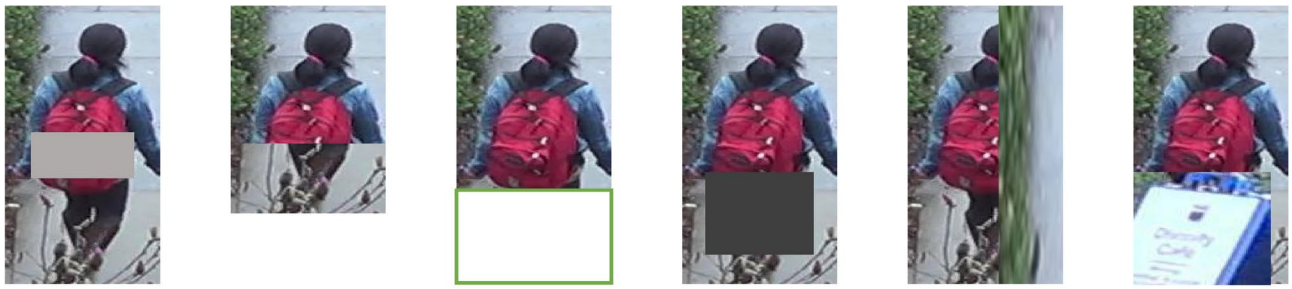

Fig. 8. Data augmentation contrast effect diagram. The images of pedestrians in this figure are from the Occluded-DukeMTMC dataset. https://github.com/lightas/Occluded-DukeMTMC-Dataset.

epochs. During inference, all local and global features output by the network are concatenated along the channel dimension to form a one-dimensional feature vector, representing the pedestrian re-identification (Table 1).

In order to conduct an objective evaluation of the proposed method, a comparative analysis is performed with state-of-the-art occluded pedestrian re-identification techniques using the extensive Occluded-DukeMTMC dataset https://github.com/lightas/Occluded-DukeMTMC-Dataset, the Occluded-ReID dataset https://cs.paperswithcode.com/dataset/occluded-reid, as well as the well-established Market-1501 https://github.com/sybernix/market1501 and DukeMTMC-ReID https://drive.google.com/file/d/1jjE85dRCMOgRtvJ5RQV9-Afs-2_5dY3O/view pedestrian re-identification datasets.

Analysis of AIRHF-net effectiveness

Effectiveness of ODAS

The experimental results comparing indices 1 and 2 in Table 2 demonstrate that the incorporation of ODAS into the baseline network improves the rank-1 accuracy by 1.7% and the mAP by 2.5%. This indicates that introducing an occlusion sample augmentation strategy allows the model to better adapt to various occlusion scenarios that may occur in the real world. By including occlusions in the training data, the model can learn more robust features, thereby enhancing its ability to handle occlusions during testing and ultimately improving its performance in practical applications. Figure 8 illustrates the comparison of the proposed algorithm with other data augmentation methods. From the results, it can be seen that the proposed ODAS is closer to real-world scenarios, which is very positive for enhancing algorithm performance.

In our experiments, we have included a comparative analysis of various data augmentation strategies, specifically comparing RE, VPM, AP-Net, IGOAS, and OAMN (Fig. 9). The experimental results indicate

that our proposed method outperforms the other methods, showing a significant performance advantage. Specifically, our method achieved high scores of 61.6% and 69.6% at two metrics, which are the highest among all the compared methods. This not only highlights the superior performance and reliability of our method for this task but also reflects its effectiveness. The primary reason for this is that, from a practical perspective, occlusions most commonly occur at the legs, sides, and above pedestrians. This suggests that our method is more closely aligned with real-world scenarios, thereby enhancing its overall performance.

The ablation study on data augmentation, as shown in Table 1, compares random occlusion, horizontal pedestrian occlusion, vertical pedestrian occlusion, and the proposed method of occluding from top, bottom, left, and right. The experimental results indicate that the horizontal occlusion method achieves a rank-1 accuracy and mAP of 69.1% and 61.3%, respectively, which are 0.5% and 0.3% lower than those of the horizontal occlusion method. Furthermore, the proposed method, which involves pasting to create occlusions from all four sides, improves the performance over the baseline by 2.5% in terms of rank-1 accuracy. This demonstrates the effectiveness of the proposed MASK strategy. The primary reason for this is that, from a practical perspective, occlusions often occur at the legs, sides, and above pedestrians, indicating that our method is more closely aligned with real-world scenarios, thereby enhancing the overall performance.

Effectiveness of AL-WIE

The AL-WIE module proposed in this paper aims to reduce the computational complexity of ViTs while effectively modeling more discriminative global features. There are various approaches to reduce computational complexity: the most straightforward solution is to decrease the embedding dimension, which may result in the loss of some detailed information. Additionally, reducing the number of attention heads is also a commonly used method; however, each head attends to different positions and features of the input sequence, which may compromise performance. In contrast, this paper adopts a different approach through the AL-WIE module by filtering out useless redundant features, thereby enhancing the capture of discriminative features. As shown in the results of Table 2, Index 3 achieves a 2.8% improvement in the rank-1 metric compared to Index 1, which uses only the baseline. Additionally, the computational complexity of AL-WIE is lower than that of a conventional Transformer encoder.This indicates that the proposed adaptive screening method can, to some extent, select more discriminative features and reduce feature redundancy, thereby lowering complexity. The comparison between Index 5 and Index 2 reveals a substantial enhancement in model performance, indicating that in the context of a large-scale occlusion dataset, the AL-WIE module can effectively improve the network's ability to perceive occlusions and construct more discriminative global feature representations.

Although sparse attention mechanisms can enhance computational efficiency, their advantages may diminish when processing long sequences. Long sequences often carry complex contextual information, some of which may be essential for understanding the content. Under sparse mechanisms, the model may fail to adequately capture all critical contextual information within long sequences, resulting in significant performance degradation. To verify that the proposed AL-WIE can further address this issue under sparsification, we compare the performance of our method with the classic sparsification strategy, top-K, on the Occluded-DukeMTMC Dataset. As shown in the Fig. 10, our method achieves improvements of 1.2% and 0.5% in rank1 and mAP metrics, respectively, compared to the top-K strategy. This performance enhancement demonstrates that the adaptive token selection module based on relation awareness can effectively mitigate the information loss caused by sparsification by considering both the correlation between class tokens and image tokens, as well as the correlations among image tokens.

Effectiveness of LHE

The experimental results from indices 4, 6, and 7 in Table 2indicate that the addition of the LHE module further enhances model performance. The reason for this phenomenon lies in the fact that stacking LHE allows for the further extraction of discriminative local features under the guidance of global semantics. Additionally, as the number of segmented local features increases, this approach is more beneficial for capturing diverse fine-grained cues. In contrast, traditional hard segmentation of local features undermines the spatial correlations among those local features.

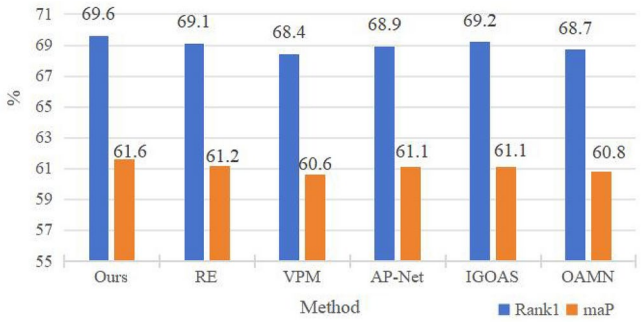

Fig. 9. The performance comparison of different data augmentation methods.

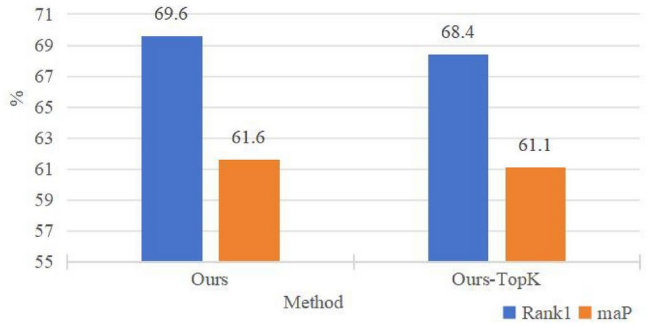

Fig. 10. The comparison of our method with the performance of the framework based on top-K sparsification.

Variants	$[S_1, S_2, S_3, S_4]$	FLOPs (G)	Rank-1	mAP
V_1	[11, 0, 0, 0]	10.84	68.1	59.8
V_2	[11, 0, 0, 0]	14.31	68.7	60.7
V_3	[11, 0, 0, 0]	16.03	68.4	60.3
V_4	[2, 9, 0, 0]	11.14	68.4	58.8
V_5	[0, 9, 2, 0]	12.45	69.6	**61.6**
V_6	[0, 9, 0, 2]	15.46	**69.7**	61.5

Table 3. Ablation experiment on the size of patches divided by tokenization operation. The best performance values are in bold.

Methods	k	Occluded-DukeMTMC		Params	FLOPs
		Rank-1	mAP	(M)	(G)
Baseline + LHE	–	60.7	52.3	87.36	17.56
	1	61.1	53.7	98.33	19.79
	2	61.9	56.3	108.67	20.93
	4	62.8	56.7	115.19	21.41
	8	**63.1**	**56.8**	121.50	22.38
	10	63.0	56.6	123.20	24.10

Table 4. Ablation experiments of LHE stacking layers. The best performance values are in bold.

Ablation experiment on image patch size from tokenization operations

In CNN architectures, shallow layers typically emphasize features with maximum spatial dimensions and minimal channel dimensions, gradually increasing channel dimensions while decreasing spatial dimensions. This design philosophy enhances the network's generalization capabilities. Consequently, we modified the tokenization operation t_1 within the AL-WIE module to yield five variants, as presented in Table 3, while maintaining a constant depth of 11 layers for the AL-WIE stack. Different tokenization operations can produce image patches of various sizes, as shown in Table 3, where S_1 indicates an image patch of size 128×128 derived from the tokenization operation t_1 , with S_1, S_2, S_3, S_4 sequentially halving this size. Notably, [11, 0, 0, 0] represents window token sizes of 128×128 across all 11 layers of AL-WIE.

The results indicate that reducing the size of the image patches increases the floating-point computational load. Moreover, decreasing the patch size in subsequent layers aids in enhancing matching accuracy. Considering a balance between computational requirements and model precision, this study concludes that V_5 represents a reasonable compromise.

Ablation experiment on the recursive stacking depth of the LHE module

In the LHE module, a recursive stacking approach is employed to delineate local features at different hierarchical levels, enabling the model to learn more optimal local features under the guidance of global semantic information. The results of the ablation experiment are shown in Table 4. As the number of LHE stacked layers increases, the rank-1 accuracy improves significantly. Furthermore, it is observed that with the enhancement in performance, both the model parameter count and floating-point computational load also increase. For example, when k = 2, the rank-1 metric improves by 1.2%, but there is a corresponding increase in both the model parameter

count and floating-point calculations compared to the Baseline. To achieve a balance between accuracy and complexity, the value for the LHE stacking depth is set to 4.

Comparative experiments

To validate the performance of the proposed model in the pedestrian re-identification task, a series of experiments were conducted. This study specifically focuses on occluded pedestrian re-identification datasets, local pedestrian re-identification datasets, and complete pedestrian re-identification datasets. A detailed experimental protocol was designed, and comparisons were made with other methods. The compared methods include both classical approaches and representative research ideas from the past three years. Overall results indicate that the proposed network achieved outstanding performance across all three datasets.

Comparison of occluded pedestrian re-identification methods

Table 5 presents the experimental comparison results of the AIRHF method against other approaches on the Occluded-DukeMTMC dataset. The methods included in the comparison are categorized into three groups: the first group employs horizontal segmentation and alignment to extract local features of pedestrians; the second group consists of methods for augmenting occluded samples; and the third group is based on mainstream network models such as ViT. As shown in the results in Table 5, the proposed AIRHF method achieved a rank-1 accuracy of 69.6% and mAP of 61.6%. Compared to existing state-of-the-art methods, the proposed approach demonstrates superior performance. In the first group, PCB and Adver Occulded effectively mined different local features through horizontal partitioning, achieving commendable performance. This further validates the rationale behind focusing on local feature learning in pedestrian re-identification research. In the proposed method, the AL-WIE constructs meaningful sequences, while the LHE progressively extracts multi-granularity local features under the guidance of global descriptors. This approach aids in better capturing local details within the images.

Compared to the second group of methods that enhance data diversity, the proposed model simulates realistic occluded scenarios by pasting actual occlusions onto the original images, ensuring that the same batch of data contains consistent occlusions to maintain data integrity, thereby attaining superior performance. When compared to DNL+BED[15] , the rank-1 accuracy and mAP are enhanced by 1.4% and 4.0%, respectively. While traditional approaches help mitigate the risk of overfitting, their generalization capability is relatively weak when confronted with diverse occlusions, which is a significant performance bottleneck. Thus, simulating more realistic occlusion scenarios is crucial for pedestrian re-identification research. Methods based on Transformer architectures are inherently designed to extract features with global relevance; however, they do not effectively utilize the local feature correlations within image sequences. The proposed method adaptively mines important information regions within images and improves the network's ability to extract context features at different scales. During the local feature extraction phase, recursive partitioning of the image sequences generates hierarchical features for distinct local traits, effectively reducing the similarity among local features and

Methods	Occluded-DukeMTMC		Occluded-REID	
	Rank-1	mAP	Rank-1	mAP
Part Aligned[40]	28.8	20.2	–	–
MGN[41]	41.2	33.4	–	–
PCB + RPP[42]	42.6	33.7	41.3	38.9
Adver Occulded[43]	44.5	32.2	–	–
IGOAS[44]	60.1	49.4	–	–
OAMN[12]	62.6	46.1	–	–
DRL-Net[15]	65.8	53.9	–	–
DNL + BED[45]	68.4	57.2	–	–
Pirt[46]	60.0	50.9	–	–
PAT[47]	64.5	53.6	81.6	72.1
PVT[48]	65.5	57.6	79.1	74.0
PAFormer[49]	66.4	60.4	–	–
TransReID[19]	66.5	57.4	84.2	78.7
AAFormer[50]	67.0	58.2	–	–
LoGoViT[51]	67.4	61.4	–	–
FED[20]	67.9	56.3	86.3	79.3
MVI2P*[52]	68.6	57.3	–	–
ViT-SPT*[53]	68.6	57.4	86.8	**81.3**
PVT*[48]	69.0	61.2	83.3	77.5
Ours	**69.6**	**61.6**	**86.8**	80.2

Table 5. Comparison with state-of-the-arts on the Occluded-DukeMTMC and Occluded-REID dataset (%). The best performance values are in bold.

significantly enhancing their distinguishability. When compared to AAFormer[50], our method improves the rank-1 metric by 2.6%, demonstrating a stronger focus on local feature learning, particularly showcasing its best performance on the Occluded-DukeMTMC dataset. Compared to the latest Transformer-based methods, MVI2P*[52], ViT-SPT*[53] , and PVT*[48], our method achieves improvements of 1%, 1%, and 0.4% respectively in the Rank-1 metric. These performance gains are primarily attributed to the synergistic contributions of the three modules proposed in this work.

Comparative experiments of local pedestrian re-identification Methods

To further evaluate the effectiveness of the proposed AIRHF method, comparative experiments were conducted on two local datasets. The Partial-ReID dataset is specifically designed to investigate the problem of local pedestrian re-identification, where pedestrian images are divided into multiple local regions, each of which may be subject to occlusion. The Partial-ilIDS dataset aims to address pedestrian re-identification under conditions where local regions exhibit occlusions due to obstructing objects. In this context, more discriminative local features are essential for local pedestrian re-identification tasks.

Methods based on CNNs and ViTs often struggle to maintain optimal accuracy across local datasets. In contrast, our proposed method takes into account the issue of uneven sample distribution by designing an ODSA process to ensure the quality and diversity of generated samples. Additionally, the AL-WIE and LHE modules implemented in our method place greater emphasis on the partitioning of local regions and the robustness of local features.

Due to the limited number of images in the aforementioned datasets, Market-1501 was utilized as the training dataset for evaluation. As presented in Table 6, the rank-1 accuracy of AIRHF on the Partial-ReID dataset reached 87.3%. Notably, our method even outperformed those utilizing external cues. In summary, the proposed model demonstrates exceptional performance on local pedestrian re-identification datasets.

Comparative experiments of complete pedestrian re-identification methods

In complete pedestrian re-identification datasets, the considerations of occlusions or other local information deficits are typically absent; therefore, it is crucial to evaluate the proposed method on non-occluded datasets. Table 7 presents the performance of AIRHF on the Market-1501 and DukeMTMC datasets. The methods compared can be categorized into four groups: methods based on global features (SFT[58], Circle[59], and IANet[60]), methods based on local features (PCB[42], AWPCN[61], VPM[62]), methods relying on external cues (Pirt[46], HOReID[57], PGFA[63]), and methods based on Transformer architectures (AAFormer[50], FCFormer[16], ViT-SPT*[53] MVI2P*[52]).

When compared to IANet from the first group, the proposed method demonstrates improvements of 1.0% in rank-1 and 5.6% in mAP metrics on the Market-1501 dataset. Furthermore, the performance surpasses that of the advanced method AWPCN from the second group by a substantial margin. This indicates that hard partitioning may not perform well when handling occlusions and is sensitive to pose variations; if part of a pedestrian is occluded, the partitioned region may fail to encompass vital information pertaining to the individual.

The LHE designed in this study performs recursive partitioning of the image sequences, allowing for different levels of local features to represent distinct semantic regions. Guided by global semantic information, it extracts more discriminative local features, in contrast to the traditional hard partitioning approach.

In comparison with the externally cue-dependent HOReID method, our approach achieves improvements of 1.2% in rank-1 and 3.8% in mAP, signifying that it achieves superior performance without relying on external semantic information or additional cues. Additionally, when compared to Transformer-based methods, AIRHF exhibits competitive results.

Complexity analysis

The ViT has demonstrated significant success in pedestrian re-identification due to its capacity to model long-range dependencies and effectively construct global feature information. However, its model parameters and computational complexity are substantially greater than those of mainstream Convolutional Neural Networks

Methods	Partial-ReID		Partial-iLIDS	
	Rank-1	mAP	Rank-1	mAP
Part Bilinear[54]	57.7	59.3	–	–
PCB+RPP[42]	66.3	63.8	–	–
MHSA-Net[55]	81.3	–	73.6	85.4
FRR[56]	81.0	76.6	–	–
PFT[39]	81.3	79.9	74.8	87.3
FED[20]	83.1	80.5	–	–
HOReID[57]	85.3	–	72.6	86.4
OAMN[12]	86.0	–	–	–
Ours	**87.3**	**80.6**	**79.8**	**87.4**

Table 6. Comparison with the state-of-the-art methods on the Partial-ReID and Partial-iLIDS datasets. The best performance values are in bold.

Methods	Market-1501		DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
SFT[58]	93.4	82.7	86.9	73.2
Circle[59]	94.2	84.9	–	–
IANet[60]	94.4	83.1	87.1	73.4
PCB[42]	92.3	77.4	81.8	66.1
AWPCN[61]	94.0	82.1	85.7	74.1
VPM[62]	93.0	80.8	83.6	72.6
Pirt[46]	94.1	86.3	88.9	77.6
HOReID[57]	94.2	84.9	86.9	75.6
PGFA[63]	91.2	76.8	82.6	65.5
ViT-SPT*[53]	94.5	86.2	89.4	79.1
PAT[47]	95.4	88.0	88.8	78.2
DRL-Net[15]	94.7	86.9	88.1	76.6
FCFormer[16]	95.0	86.8	89.7	78.8
MVI2P*[52]	95.3	87.9	–	–
AAFormer[50]	95.4	87.7	90.1	80.0
Ours	95.4	88.7	90.5	81.7

Table 7. Comparison with the state-of-the-art methods on the Market-1501 and DukeMTMC-ReID datasets.

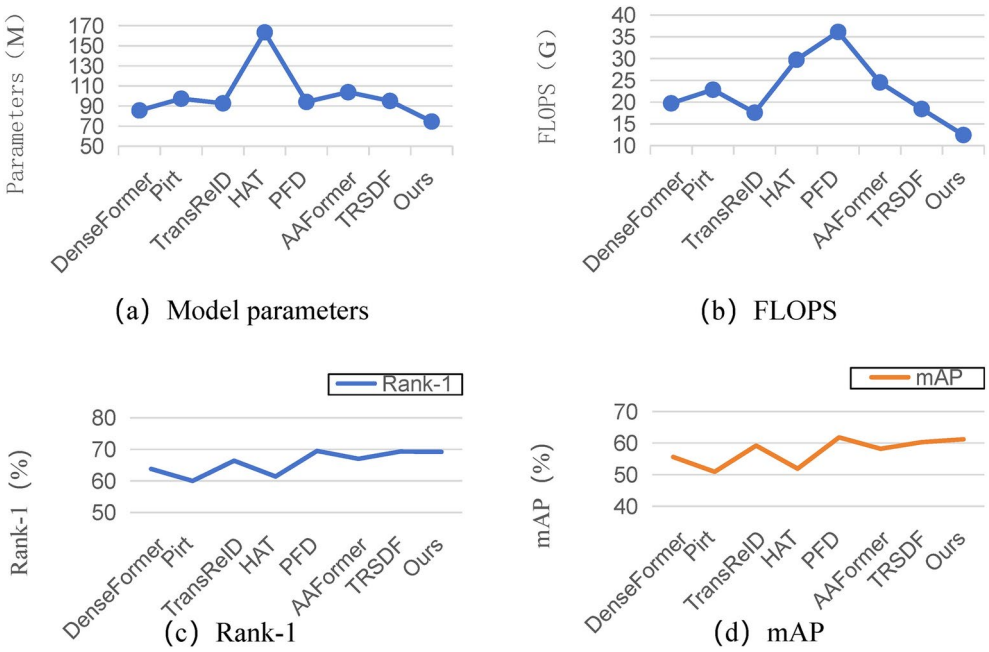

Fig. 11. Comparison of model complexity on the Occluded-DukeMTMC dataset.

(CNNs). Therefore, designing a lightweight re-identification model is essential. This study compares the proposed method's complexity and accuracy with prior research based on ViT, including DenseFormer[64], Pirt[46], TransReID[19] , HAT[5], PFD[65], AAFormer[50], and TRSDF[66].

To ensure fair experimental results, all ViT implementations were set to the base type. As illustrated in Fig. 11, the comprehensive performance of AIRHF surpasses that of the other methods. Specifically, the proposed method achieves a floating-point computation volume that is 54% and 51% of that of Pirt and AAFormer, respectively, while improving rank-1 accuracy by at least 1.5%. Although it exhibits slightly lower performance than PFD, the proposed method reduces both the parameter count and floating-point computation volume compared to PFD by 21% and 66%, respectively.

In contrast to overly subjective manual design strategies for attention sparsification, our algorithm introduces AL-WIE, which decreases computational overhead while simultaneously mitigating the network's focus on occluded regions. In summary, our method not only achieves outstanding matching accuracy but also demonstrates strong advantages in terms of complexity.

Visualization analysis

In this section, we analyze the proposed method from two perspectives: attention mechanism visualization and ranking results visualization.

1. Attention Heatmap Visualization: As shown in Fig. 12, (a) displays the original image, (b) represents the heatmap based on the Vision Transformer, and (c) illustrates the heatmap generated by the proposed AIRHF method. This section presents a comparison of the attention heatmaps between the proposed AIRHF algorithm and the baseline network, the Vision Transformer model. The AIRHF method, by considering the importance of image patches and utilizing aLHE, demonstrates a significant improvement in extracting discernible features of pedestrians compared to the traditional Vision Transformer model, particularly when it comes to items like backpacks. Additionally, it effectively reduces the network's focus on occluded information.
2. Visualization of Pedestrian Image Retrieval Results Ranking: The ranking results of the proposed method are illustrated in Fig. 13, where the far-left image represents the query image, and the right side displays the 10 closest matching results. Green indicates correct retrieval results, while red signifies incorrect ones. The results demonstrate a high level of accuracy in the retrieval outcomes. From the results, it is clear that the overall performance of the algorithm is quite good; however, it occasionally makes errors when faced with complex backgrounds or when pedestrians have a high degree of similarity. These errors are likely attributed to feature loss during the processes of sparsification and feature selection. In summary, although the proposed method performs well under typical conditions, the inherent trade-off between computational efficiency and feature retention signifies some degree of information loss. This may lead to occasional misidentifications under challenging conditions. Future work can focus on alleviating these issues to further enhance the robustness of the system.

Conclusion

This paper presents a local representation learning algorithm based on feature fusion. The algorithm includes the ODAS, AL-WIE, and LHE modules. The ODAS module constructs a set of occlusions to ensure that they do not appear in the test set, thereby augmenting the training dataset by pasting them onto the original images, which enhances the network model's ability to perceive occluded information.

Furthermore, compared to previous sparse attention approaches, the proposed AL-WIE is more suited for the research of occluded pedestrian re-identification. As the network training progresses through iterations, the attention interaction gradually focuses on crucial image patches, achieving a balance between performance and computational cost. Additionally, AL-WIE incorporates multi-scale information into image patches of varying sizes, enabling it to adaptively capture different scale information present in the input images.

The LHE module is subsequently introduced to extract spatial correlation features of sequences under the guidance of global semantics, while hierarchical feature learning is employed to mine discriminative local information. Finally, comprehensive experiments were conducted on large-scale datasets, demonstrating that the proposed model relies on a relatively low amount of computation and trainable parameters, providing insights and references for applications with high real-time requirements.

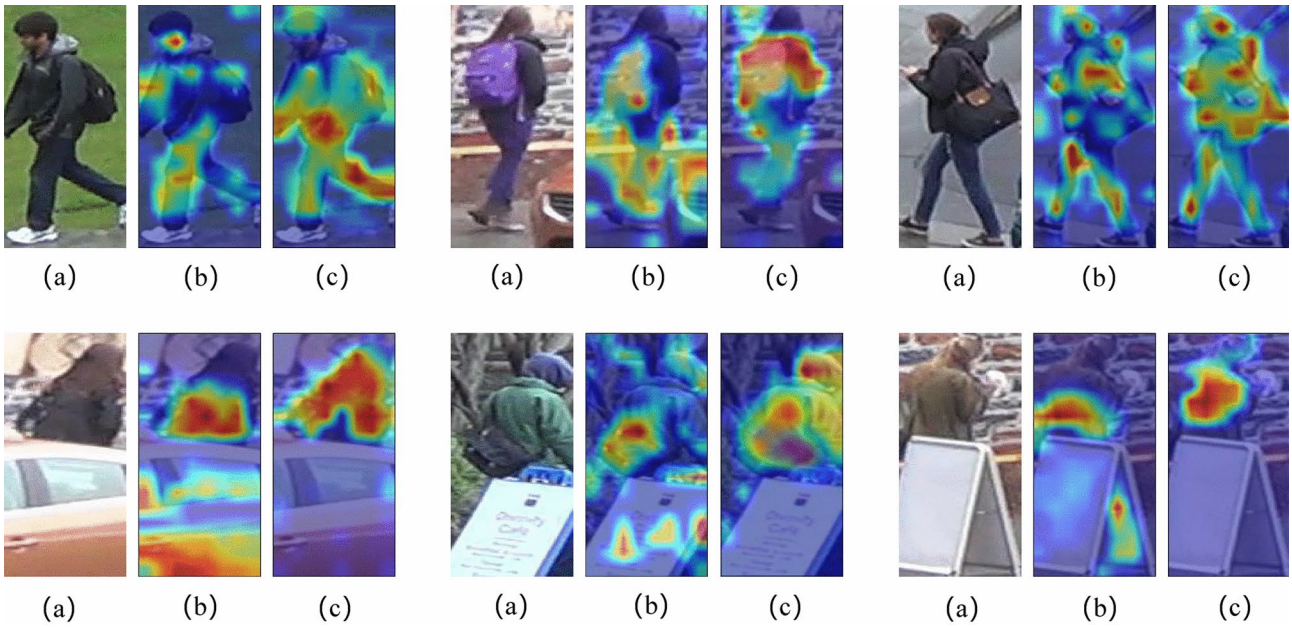

Fig. 12. Attention heatmaps based on different Vision Transformer model.The images of pedestrians in this figure are from the Occluded-DukeMTMC dataset. https://github.com/lightas/Occluded-DukeMTMC-Dataset.

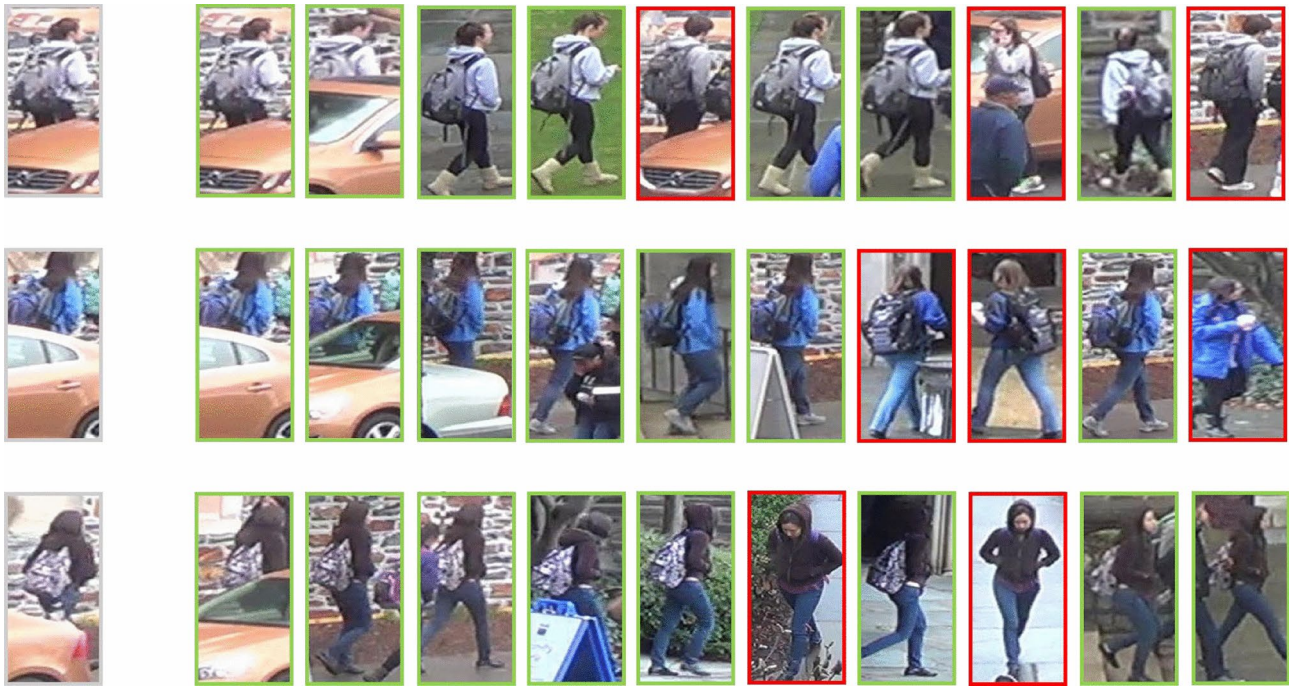

Fig. 13. Schematic diagram of sorting results. The images of pedestrians in this figure are from the Occluded-DukeMTMC dataset. https://github.com/lightas/Occluded-DukeMTMC-Dataset.

Data availability

The datasets generated and/or analyzed during the current study are available in this repository. https://github.com/ggsszz123/Person-reid-Dataset . Specific contents include: Occluded-DukeMTMC: https://github.com/lightas/Occluded-DukeMTMC-Dataset. Partial-ReID:url:JDAI-CV/Partial-Person-ReID(github.com) Partial-iLIDS: https://drive.google.com/file/d/1ErCEQsNHSHpgZF3-NNj6_OH322vpk8gn/view?usp=sharing Market-1501: https://github.com/sybernix/market1501 DukeMTMC:https://drive.google.com/file/d/1jjE85dRCMOgRtvJ5RQV9-Afs-2_5dY3O/view.

Received: 28 August 2024; Accepted: 16 October 2024
Published online: 08 November 2024

References

1. Yan, G., Wang, Z., Geng, S., Yu, Y. & Guo, Y. Part-based representation enhancement for occluded person re-identification. *IEEE Trans. Circ. Syst. Video Technol.* **33**, 4217–4231 (2023).
2. Ning, E., Wang, Y., Wang, C., Zhang, H. & Ning, X. Enhancement, integration, expansion: Activating representation of detailed features for occluded person re-identification. *Neural Netw.* **169**, 532–541 (2024).
3. Nguyen, V. D. et al. Tackling domain shifts in person re-identification: A survey and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4149–4159 (2024).
4. Akan, S., Varlı, S. & Bhuiyan, M. A. N. An enhanced swin transformer for soccer player reidentification. *Sci. Rep.* **14**, 1139 (2024).
5. Zhang, G., Zhang, P., Qi, J. & Lu, H. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 516–525 (2021).
6. Zhang, Y. et al. Local correlation ensemble with gcn based on attention features for cross-domain person re-id. *ACM Trans. Multimed. Comput. Commun. Appl.* **19**, 1–22 (2023).
7. Sarker, P. K., Zhao, Q. & Uddin, M. K. Transformer-based person re-identification: A comprehensive review. *IEEE Trans. Intell. Veh.* **2024**, 59 (2024).
8. Bai, S., Chang, H. & Ma, B. Incorporating texture and silhouette for video-based person re-identification. *Pattern Recogn.* **156**, 110759 (2024).
9. Zhu, H., Budhwant, P., Zheng, Z. & Nevatia, R. Seas: Shape-aligned supervision for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 164–174 (2024).
10. Wang, Z., Huang, H., Zheng, A., Li, C. & He, R. Parallel augmentation and dual enhancement for occluded person re-identification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3590–3594 (IEEE, 2024).
11. Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34* 13001–13008 (2020).
12. Chen, P. et al. Occlude them all: Occlusion-aware attention network for occluded person re-id. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11833–11842 (2021).
13. Liao, K. et al. Occluded person re-id based on dual attention mask guidance. *Int. J. Multimedia Inf. Retrieval.* **2024**, 8569 (2024).
14. Li, Y. et al. Occlusion-aware transformer with second-order attention for person re-identification. *IEEE Trans. Image Process.* **2024**, 745 (2024).
15. Jia, M. et al. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Trans. Multimedia.* **25**, 1294–1305 (2022).
16. Wang, T. et al. Feature completion transformer for occluded person re-identification. *IEEE Trans. Multimedia.* **2024**, 52 (2024).

17. Wu, X. et al. Text-based occluded person re-identification via multi-granularity contrastive consistency learning. In *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38* 6162–6170 (2024).
18. Guo, X. et al. A novel dual-pooling attention module for uav vehicle re-identification. *Sci. Rep.* **14**, 2027 (2024).
19. He, S., Luo, H., Wang, P. et al. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14993–15002. https://doi.org/10.1109/ICCV48922.2021.01474 (IEEE, 2021).
20. Wang, Z., Zhu, F., Tang, S. et al. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4754–4763 (2022).
21. Wang, S. et al. Occluded person re-identification via defending against attacks from obstacles. *IEEE Trans. Inf. Forensics Secur.* **18**, 147–161. https://doi.org/10.1109/TIFS.2022.3218449 (2023).
22. Wang, Y., Li, Y. & Cui, Z. Incomplete multimodality-diffused emotion recognition. *Adv. Neural Inf. Process. Syst.* **36**, 56 (2024).
23. Wang, Y., Cui, Z. & Li, Y. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22025–22034 (2023).
24. Wang, Y., Lu, T., Yao, Y., Zhang, Y. & Xiong, Z. Learning to hallucinate face in the dark. *IEEE Trans. Multimedia* (2023).
25. Wang, Y. et al. Faceformer: Aggregating global and local representation for face hallucination. *IEEE Trans. Circ. Syst. Video Technol.* **33**, 2533–2545 (2022).
26. Song, Y. & Liu, S. A deep hierarchical feature sparse framework for occluded person re-identification. arXiv preprint arXiv:2401.07469 (2024).
27. Rao, Y. et al. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Adv. Neural Inf. Process. Syst.* **34**, 13937–13949 (2021).
28. Meng, L. et al. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 12299–12308 (IEEE, 2022).
29. Zhang, Z., Han, S., Liu, D. & Ming, D. Focus and imagine: Occlusion suppression and repairing transformer for occluded person re-identification. *Neurocomputing.* **578**, 127442 (2024).
30. Liang, Y. et al. Evit: Expediting vision transformers via token reorganizations. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 1–21 (2022).
31. Yin, H. et al. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10799–10808 (IEEE, 2022).
32. Zhuo, J., Chen, Z., Lai, J. & Wang, G. Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6 (IEEE, 2018).
33. Dong, X. et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12124–12134 (2022).
34. Fang, Y., Wang, X., Wu, R. & Liu, W. What makes for hierarchical vision transformer?. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 12714–12720 (2023).
35. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020).
36. Fournier, Q., Caron, G. M. & Aloise, D. A practical survey on faster and lighter transformers. *ACM Comput. Surv.* **55**, 1–40 (2023).
37. Cheng, K., Tang, J., Gu, H., Wan, H. & Li, M. Cross-block sparse class token contrast for weakly supervised semantic segmentation. *IEEE Trans. Circ. Syst. Video Technol.* (2024).
38. Rong, L. et al. A vehicle re-identification framework based on the improved multi-branch feature fusion network. *Sci. Rep.* **11**, 20210 (2021).
39. He, L., Wang, Y., Liu, W. et al. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8450–8459 (2019).
40. Zhao, L., Li, X., Zhuang, Y. & Wang, J. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision* 3219–3228 (2017).
41. Wang, G., Yuan, Y., Chen, X., Li, J. & Zhou, X. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM International Conference on Multimedia*. 274–282 (2018).
42. Sun, Y., Zheng, L., Yang, Y., Tian, Q. & Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*. 480–496 (2018).
43. Huang, H., Li, D., Zhang, Z. et al. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5098–5107 (IEEE, 2018).
44. Zhao, C. et al. Incremental generative occlusion adversarial suppression network for person reid. *IEEE Trans. Image Process.* **30**, 4212–4224 (2021).
45. Yan, C., Pang, G., Jiao, J. et al. Occluded person re-identification with single-scale global representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11875–11884 (2021).
46. Ma, Z., Zhao, Y. & Li, J. Pose-guided inter-and intra-part relational transformer for occluded person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1487–1496 (2021).
47. Li, Y., He, J., Zhang, T. et al. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2898–2907 (2021).
48. Gao, H. et al. Point-level feature learning based on vision transformer for occluded person re-identification. *Image Vis. Comput.* **143**, 104929 (2024).
49. Jung, H., Lee, J., Yoo, J., Ko, D. & Kim, G. Paformer: Part aware transformer for person re-identification. arXiv preprint arXiv:2408.05918 (2024).
50. Zhu, K., Guo, H., Zhang, S. et al. Aaformer: Auto-aligned transformer for person re-identification. *IEEE Trans. Neural Netw. Learn. Syst.* (2023).
51. Nguyen Phan, T. D. H. et al. Logovit: Local-global vision transformer for object re-identification. In *ICASSP*. 1–5 (2023).
52. Dong, N., Yan, S., Tang, H., Tang, J. & Zhang, L. Multi-view information integration and propagation for occluded person re-identification. *Inf. Fusion* (2024).
53. Tan, L. et al. Occluded person re-identification via saliency-guided patch transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38*. 5070–5078 (2024).
54. Suh, Y., Wang, J., Tang, S. et al. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 402–419 (2018).
55. Tan, H., Liu, X., Yin, B. et al. Mhsa-net: Multihead self-attention network for occluded person re-identification. *IEEE Trans. Neural Netw. Learn. Syst.* (2022).
56. Zhao, Y. et al. Short range correlation transformer for occluded person re-identification. *Neural Comput. Appl.* **34**, 17633–17645 (2022).
57. Wang, G., Yang, S., Liu, H. et al. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6449–6458 (2020).
58. Luo, C., Chen, Y., Wang, N. et al. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4976–4985 (2019).
59. Sun, Y., Cheng, C., Zhang, Y. et al. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6398–6407 (2020).
60. Hou, R., Ma, B., Chang, H. et al. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9317–9326 (2019).

61. Shu, X. et al. Adaptive weight part-based convolutional network for person re-identification. *Multimedia Tools Appl.* **79**, 23617–23632 (2020).
62. Sun, Y., Xu, Q., Li, Y. et al. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 393–402 (2019).
63. Miao, J., Wu, Y., Liu, P. et al. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 542–551 (2019).
64. Ma, H. et al. Denseformer: A dense transformer framework for person re-identification. *IET Comput. Vision.* **17**, 527–536 (2023).
65. Wang, T., Liu, H., Song, P. et al. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 2540–2549. https://doi.org/10.1609/aaai.v36i3.20155 (2022).
66. Yan, G., Wang, H. & Geng, S. Token recombination based shallow-deep feature fusion method for occluded person re-identification. arXiv preprint 1–21 (2024).

Acknowledgements

Tianjin Municipal Education Commission Research Plan Project(No.2022KJ110).

Author contributions

S. G is responsible for code writing, experimental setup, and writing; H.W. W. is responsible for experimental design, Q.Y.is responsible for method design and writing, and Z.Y.S. is responsible for method design and code writing. All authors reviewed the manuscript.

Competing interests

We declare that we have no conflicts of interest regarding the research conducted in this study. All authors contributed to the research independently, and there are no financial or personal relationships with other people or organizations that could inappropriately influence our work. As academic researchers, our primary aim is to uphold the integrity and objectivity of our research endeavors.

Additional information

Correspondence and requests for materials should be addressed to Q.Y.