Complex & Intelligent Systems (2024) 10:4557–4569
https://doi.org/10.1007/s40747-024-01395-2

ORIGINAL ARTICLE

Discriminative multi-scale adjacent feature for person re-identification

Mengzan Qi[1] · Sixian Chan[1] · Feng Hong[2] · Yuan Yao[3] · Xiaolong Zhou[4]

Received: 9 September 2023 / Accepted: 16 February 2024 / Published online: 21 March 2024

Abstract

Recently, discriminative and robust identification information has played an increasingly critical role in Person Re-identification (Re-ID). It is a fact that the existing part-based methods demonstrate strong performance in the extraction of fine-grained features. However, their intensive partitions lead to semantic information ambiguity and background interference. Meanwhile, we observe that the body with different structural proportions. Hence, we assume that aggregation with the multi-scale adjacent features can effectively alleviate the above issues. In this paper, we propose a novel Discriminative Multi-scale Adjacent Feature (MSAF) learning framework to enrich semantic information and disregard background. In summary, we establish multi-scale interaction in two stages: the feature extraction stage and the feature aggregation stage. Firstly, a Multi-scale Feature Extraction (MFE) module is designed by combining CNN and Transformer structure to obtain the discriminative specific feature, as the basis for the feature aggregation stage. Secondly, a Jointly Part-based Feature Aggregation (JPFA) mechanism is revealed to implement adjacent feature aggregation with diverse scales. The JPFA contains Same-scale Feature Correlation (SFC) and Cross-scale Feature Correlation (CFC) sub-modules. Finally, to verify the effectiveness of the proposed method, extensive experiments are performed on the common datasets of Market-1501, CUHK03-NP, DukeMTMC, and MSMT17. The experimental results achieve better performance than many state-of-the-art methods.

Keywords Person re-identification · Feature extraction · Feature aggregation · Discriminative feature

Introduction

With the rapid advancement of surveillance technologies and the exponential growth of video data, there is an increasing need to accurately and efficiently identify individuals across multiple cameras and locations. As a result, the task of person re-identification (Re-ID) [1–3] emerged, providing the capability to identify the same individual from diverse databases captured by different cameras. Specifically, the person Re-ID technology assists in the identification and tracking of suspicious or wanted individuals in crowded public spaces, airports, transportation hubs, and other high-security areas. By accurately matching individuals in real-time, security personnel can rapidly respond to potential threats and take appropriate measures to ensure public safety. Overall, the person Re-ID technology plays a critical role. However, the person Re-ID task faces major challenges, including camera parameters, lighting conditions, viewpoint variations, pedestrian posture diversities, and so on.

Studies [4–6] have highlighted the importance of extracting discriminative features in the context of the person Re-ID task. However, solely focusing on extracting global-level discriminative features may lack the supervision required for capturing fine-grained features. Furthermore, within the same image, there is variation in the proportion of different body parts, such as the head and legs. Additionally, there is an inconsistency in the representation of pedestrians captured by multi-camera systems with different viewpoints. To address these challenges, researchers have proposed part-based methods [7, 8] that utilize a general structural partitioning approach. This involves splitting the feature map

✉ Sixian Chan
sxchan@zjut.edu.cn

✉ Feng Hong
hongfeng@zjsru.edu.cn

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, Zhejiang, China

[2] School of Information Science and Technology, Zhejiang Shuren University, Hangzhou 3100 00, Zhejiang, China

[3] School of Computer Science, University of Nottingham Ningbo China, Ningbo, China

[4] College of Electrical and Information Engineering, Quzhou University, Quzhou, China

into several horizontal blocks to extract limb information for each part. Although this slicing scheme can provide richer fine-grained local features, it suffers from the fragmentation of contextual information in adjacent blocks, including blocks that may contain mostly background information. Moreover, increasing the number of partitions leads to a lack of semantic information within each block and network redundancy due to excessive partitioning. Given these limitations, it is promising to explore effective approaches that facilitate interactions between adjacent feature blocks to mitigate the loss of semantic information.

On the other hand, these part-based methods [2, 4] adopt Convolutional Neural Networks (CNNs) [9, 10] for generic feature extraction. These methods employ partition schemes in the high-level and abstract representation space, which results in the lack of informative cues at the low-level features. Nevertheless, low-level features play a critical role in discrimination tasks, e.g., clothes, backpacks, or even shoes. In general, low-level features can provide more specific information, while high-level features contain more semantic information with less specific information. Therefore, it is beneficial to exploit the complementary strengths of different-level features in an effective way. However, simple fusion operations deteriorate aggregation performance, e.g., addition operation or concatenation operation. Recently, Transformer [11–14] originating from Natural Language Processing (NLP), has shown excellent performance in computer vision tasks. With its self-attention mechanism, Transformer [15, 16] effectively captures long-range dependencies among image tokens and maintains a global perspective. In contrast, CNNs suffer from a limited receptive field due to the inherent nature of convolutional operations. To address this limitation, a promising approach involves integrating CNNs and Transformers, leveraging the respective strengths of both architectures to extract discriminative information from features at different levels.

In this paper, we propose a novel Discriminative Multi-scale Adjacent Feature (MSAF) learning framework to address semantic information ambiguity and background interference due to intensive partition. It can further extract discriminative identification information and implement adjacent feature interactions with diverse scales. To obtain discriminative information from different-level features, we design a Multi-scale Feature Extraction (MFE) module by combining CNN and Transformer structure, and it contains in terms of intra-scale and inter-scale. Moreover, we propose a Jointly Part-based Feature Aggregation (JPFA) approach. As shown in Fig. 1, to adapt to the different proportions of the limbs within the same image and the inconsistent percentages of pedestrians within different viewpoints, we adopt a diverse division scheme to enhance the robustness of identification. Specifically, the Same-scale Feature Correlation (SFC) sub-module is proposed to link adjacent blocks within the same slicing scheme. The Cross-scale Feature Correlation (CFC) sub-module is designed to adapt body parts' proportions under different scales. To further aggregate diverse adjacent features in CFC, we adopt the Graph Attention (GAT) mechanism [17] to learn discriminative information. In this way, the JPFA can effectively enrich semantic information and disregard background noise. Combined with MFE and JPFA, the identity representation can be more discriminative, thereby enhancing the retrieval performance in the person Re-ID task. The main contributions of this paper can be summarized as follows:

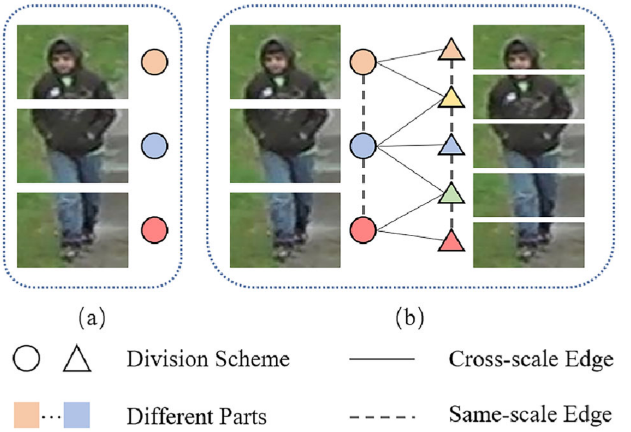

Fig. 1 Division Scheme. **a** Existing part-based works concentrate on the information within the same-scale partition scheme and ignore the associations between multi-scale adjacent feature blocks. **b** The association of adjacent feature blocks in our framework. The solid line means the cross-scale edge in Cross-scale Feature Correlation (CFC), and the dotted line indicates the same-scale edge in Same-scale Feature Correlation (SFC)

- The Multi-scale Feature Extraction (MFE) module is designed by combining CNN and Transformer structure to obtain the discriminative specific information from different-level features.
- The Jointly Part-based Feature Aggregation (JPFA) mechanism is revealed to implement adjacent feature interactions with diverse scales, which contains Same-scale Feature Correlation (SFC) and Cross-scale Feature Correlation (CFC) sub-modules. Furthermore, it can effectively alleviate semantic information ambiguity and background interference.
- Extensive experiment results on the open public challenging datasets illustrate the proposed method achieves better results than many state-of-the-art methods, e.g., Market-1501, CUHK03-NP, DukeMTMC, and MSMT17.

Related work

Part-based feature for person Re-ID

Nowadays, there are various ways [7, 18, 19] to extract part-based features on person Re-ID tasks. These methods involve dividing the feature map into horizontal blocks to extract limb information for each part. For example, Sun et al. [7] proposes a Part-Based Convolutional Baseline (PCB) network, which employs a uniform segmentation strategy with a CNN structure to capture local detail information. Considering the intensive partitioning, which can lead to ambiguity in semantic information and interference from the background, some researchers draw attention to exploring the integration of both local and global features. Towards this, Luo et al. [18] propose AlignedReID++, which learned both local and global features while dynamically aligning local information. Similarly, to incorporate richer multi-scale information, Wang et al. [4] introduces the Multiple Granularity Network (MGN), which consists of multiple global branches and various local branches for capturing multi-granularity feature representations.

However, these existing methods have a limitation in terms of aggregating multi-scale features effectively. For instance, MGN [4] relies on using a varying number of stripes in each local branch to achieve the desired multi-granularity. However, this approach treats the feature representation of each branch independently, overlooking potential interactions and correlations across different scales. As a result, there is a risk of suboptimal utilization of multi-scale information, leading to the lack of identity discrimination in the feature representations. Differing from these, we propose a novel Discriminative Multi-scale Adjacent Feature (MSAF) learning framework to extract discriminative identification information by implementing adjacent feature interactions across diverse scales. By considering the relationships between adjacent features at different scales, our approach aims to capture more comprehensive and contextually rich representations.

Feature aggregation for person Re-ID

Feature aggregation plays a crucial role in computer vision tasks, as it aims to capture both detailed and semantic information [20]. Several studies in the field of person Re-ID have attempted to incorporate feature aggregation within the network [16, 20–22]. For instance, Chen et al. [21] utilizes cascaded aggregated features to extract salient identification information. Liu et al. [22] designs a Hierarchical Bi-directional Feature Perception Network (HBFP-Net), which integrates relatively low-level features into subsequent levels. Besides, some researchers design various pyramidal structures [23, 24] to aggregate multi-scale features. These structures capture discriminative information from distinct spatial scales, with each layer focusing on specific semantic levels. Zheng et al. [24] proposes a coarse-to-fine pyramid structure for accurate bounding box predictions, while Martinel et al. [23] highlights the ability of pyramid structures to extract multi-scale representations and decompose them into distinguishing features at different semantic levels.

In recent years, Transformer [11] has shown excellent performance in capturing long-range dependencies, which provides new opportunities for feature aggregation. Inspired by this, Zhang et al. [16] leverage the Transformer framework to aggregate multiple features, enabling enhanced details while preserving semantic information. In our work, we draw inspiration from the effective feature aggregation capabilities demonstrated in these studies. Specifically, we propose a novel approach that combines CNN and Transformer structures to obtain discriminative and specific features from different levels.

Attention mechanism for person Re-ID

To emphasize important features and suppress irrelevant ones, some studies [25–28] proposed to introduce an attention mechanism in the network. Especially, in Re-ID [29–31], leveraging the relationship between different limb parts is beneficial for person feature extraction. For example, Zhang et al. [29] introduces the Relation-Aware Global Attention (RGA) model to learn the discriminative representation, which incorporates both spatial attention and channel attention. Moreover, Huang et al. [1] proposes a Three-Dimensional Transmissible Attention Network to learn identity-related information in a three-dimensional perspective. Zhu et al. [14] describes the concept of part tokens with Transformer to automatically locate limb and non-limb parts. Furthermore, He et al. [32] integrates a sequence of image patches with non-visual clues and employs the Transformer to extract robust features. In this paper, we focus on the utilization of multi-scale adjacent features to facilitate the extraction of discriminative identification information. Specifically, we explore the GAT [17] mechanism to further aggregate adjacent features across diverse scales.

Proposed method

In this section, we introduce the framework of Discriminative Multi-scale Adjacent Feature (MSAF), as shown in Fig. 2. Overall, our proposed MSAF (Multi-scale Feature Aggregation) framework comprises two main modules: the Multi-scale Feature Extraction (MFE) module and the Jointly Part-based Feature Aggregation (JPFA) module.

The MFE (left part in Fig. 2) module combines the architectures of Convolutional Neural Networks (CNNs) and

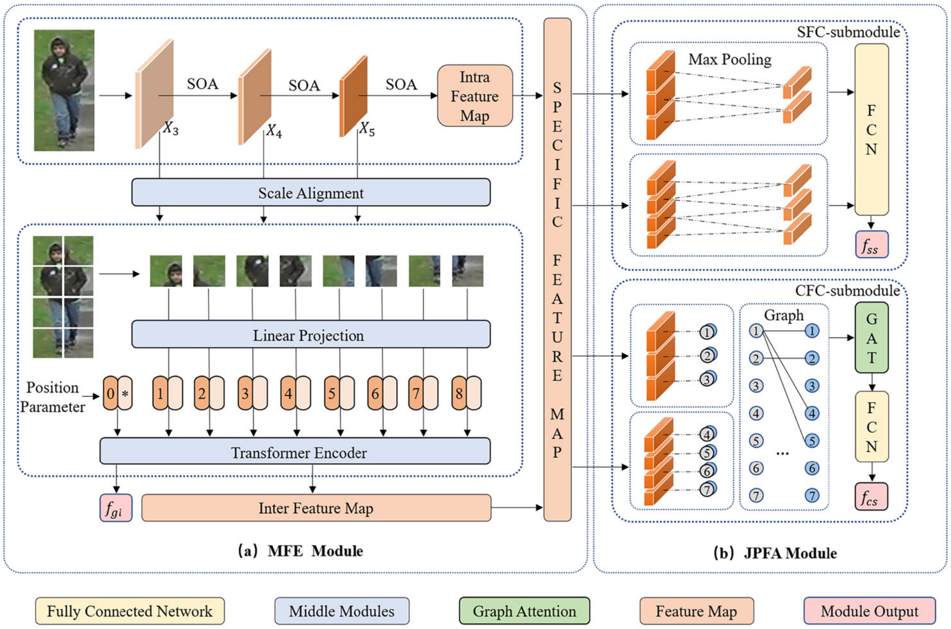

Fig. 2 The illustration of our proposed Discriminative Multi-scale Adjacent Feature (MSAF) network. Among them, the ResNet50 is utilized as the backbone. In MFE, the SOA [33] and Transformer structure [12] are designed to implement intra-scale and inter-scale feature extraction respectively. To align the size of the feature map extracted from different-layer, the Scale Alignment module simply contains the combination of Bottleneck and MaxPooling. In JPFA, we integrate SFC and CFC submodules to aggregate adjacent information from Specific Feature Map. Finally, Module Output is used to supervise the network with a multi-loss function in the training stage, while denoting identification information for each image in the test stage

Transformers to extract discriminative and specific features. Specifically, this module operates in two ways: intra-scale and inter-scale. For intra-scale, we adopt Second Order Attention (SOA) block mines contextual relevance to highlight salient features. This allows the model to focus on the fine-grained characteristics. On the other hand, the inter-scale process is implemented by the manner of Transformer, which captures the contextual and holistic information by considering the relationship between different scales. This enables the model to understand the global structure and appearance of the person.

To facilitate effective interactions between adjacent features with diverse scales, we introduce two sub-modules within the JPFA (right part in Fig. 2) module: the Same-scale Feature Correlation (SFC) and the Cross-scale Feature Correlation (CFC) sub-modules. The SFC sub-module enables the model to capture correlations and dependencies between features within the same scale. This benefits in preserving the detailed information and local characteristics within each scale. The CFC sub-module, on the other hand, facilitates the exploration of correlations and interactions between features across different scales. This allows the model to capture the relationships between different body parts across various scales.

By incorporating the MFE and JPFA modules within our MSAF framework, we aim to obtain discriminative feature representation that combines both fine-grained details and global contextual information.

Multi-scale feature extraction

Considering the powerful feature representation of CNN, we adopt the ResNet50 [34] as the backbone. In general, we obtain the feature map $X_l \in \mathbb{R}^{C_l \times H_l \times W_l}$ from l-th layer of ResNet50, where C_l, H_l and W_l respectively indicate the number of channels, the height, and width of the feature map. Firstly, we simply introduce an intra-scale feature extraction scheme. Inspired by [33], the Second Order Attention (SOA) block mines contextual relevance to highlight salient features. To take advantage of this, the SOA module is adopted to enable intra-scale alignment for feature map X_l.

$$X_l = SOA(X_l). \tag{1}$$

where SOA is the self-attention block proposed in [33]. After the intra-scale extraction within hierarchical layers, we obtain Intra Feature Map $X_{tra} \in \mathbb{R}^{C_5 \times H_5 \times W_5}$.

For inter-scale extraction in MFE, we define hierarchical feature maps from different scales as $X_1, X_2, \ldots, X_n$. To achieve consistent spatial dimensions for different scales, we employ a Scale Alignment module simply consisting of Bottleneck and MaxPooling layer. Bottleneck [34] involves a series of stacked residual blocks. And the MaxPooling operation reduces all feature maps to the same spatial size, denoted as $h \times w$.

$$X_j = MaxPooling(Bottleneck(X_j)). \tag{2}$$

where Bottleneck is originated from [34]. After that, we concatenate the hierarchical features.

$$F = Concat(X_1, X_2, ...X_n). \tag{3}$$

In this way, we obtain the hierarchical feature $F \in \mathbb{R}^{c \times h \times w}$, where $c = \sum_{i=1}^{n} C_l$. Then, we utilize the Transformer structure to interact with information obtained in F. As shown in Fig. 2a, the feature map F is sliced into $N = (h \times w)/R^2$ image patches with each patch of $(R \times R)$ size. The image patches are then projected into D-dimensional vector sequences by the linear projection layer. On meanwhile, the learnable parameters named class tokens are embedded to extract global features. For the purpose of enriched spatial position information, we also add the learnable position parameter in the vector sequence. After Position Embedding, we obtain the vector sequences $Z \in \mathbb{R}^{L \times D}$, where $L = N + 1$. In [12], the classic Transformer Encoder layer (TEL) contains stacked Multi-Head Self-Attention (MSA), Multi-Layer Perceptron (MLP), Layer Normalization (LN), and other blocks. The number of the Transformer Encoder layer mentioned above is set as d. For example, the output of l-th TEL is generated as follows:

$$Z^l = Transformer(Z^{l-1}). \tag{4}$$

where Z^0 is the original vector sequences after Position Embedding, $Z^l \in \mathbb{R}^{c' \times N}$. c' denotes the dimension of output from the l-th TEL. Following Eq. 4, we gain both global identification information f_{gl} and Inter Feature Map X_{ter}, where $f_{gl} \in \mathbb{R}^{c' \times 1}$ retrieved from class token and $X_{ter} \in \mathbb{R}^{c' \times h \times w}$ extracted from image patches with reshape operation. Finally, the Specific Feature Map $X_f \in \mathbb{R}^{c_f \times h \times w}$ is concatenated from Intra Feature Map X_{tra} and Inter Feature Map X_{ter}, where $c_f = C_5 + c'$. Furthermore, the Specific Feature Map X_f is forwarded into the Jointly Part-based Feature Aggregation (JPFA) module to interact with adjacent features with diverse scales, which is consistent with the structure of different proportions of the human body.

Jointly part-based feature aggregation

The Jointly Part-based Feature Aggregation (JPFA) module contains the Same-scale Feature Correlation (SFC) and the Cross-scale Feature Correlation (CFC) sub-modules, both of which are based on partition. To begin with, we introduce the slicing scheme $DIVISION = \{S_1, S_2, ...S_N\}$, where $S_d(d \in \{1, ...N\})$ indicates that horizontally dividing feature map X_f into d pieces of equal size. $\{S_{d,1}, S_{d,2}, ...S_{d,d}\}$ denotes each piece respectively.

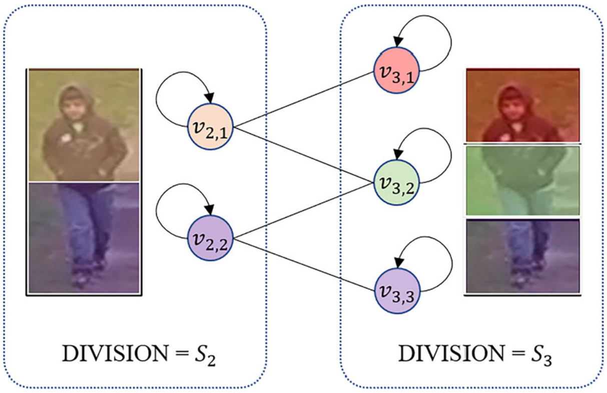

Fig. 3 The example graph with slicing scheme $DIVISION = \{S_2, S_3\}$ in CFC. The vertex set consists of feature blocks under different slicing schemes. For the pair of vertex, an undirected edge exists if their bounding boxes intersect. Specifically, each vertex has a self-loop

Same-scale feature correlation

This sub-module describes the correlation of adjacent features within the same slicing scheme. For slicing scheme S_d, we obtain d feature blocks $\{S_{d,1}, S_{d,2}, ...S_{d,d}\}$ with the same size. Firstly, we squeeze the spatial dimension of per-block by MaxPooling operation to obtain each discriminative identification, denoted as $S_{d,j}^{max} \in \mathbb{R}^{c_f \times 1 \times 1}(j \in \{1, ...d\})$. To enrich semantic information and disregard background noise, we propose the aggregation of adjacent feature. Then, we also utilize MaxPooling layer to enhance salient information for segmented adjacent blocks. It can be represented as:

$$S_{d,i}^{max\_adj} = MaxPooling(S_{d,i-1}^{max}, S_{d,i}^{max}). \tag{5}$$

where $S_{d,i-1}^{max}$ and $S_{d,i}^{max}$ represent discriminative information of adjacent blocks and $i \in \{2, ...d\}$. In this way, we obtain $S_d^{max\_adj} \in \mathbb{R}^{c_f \times (d-1)}$ for slicing scheme S_d by concatenating all the $S_{d,i}^{max\_adj}$. Finally, we extract same-scale features f_{ss}^d from $S_d^{max\_adj}$ with a fully connected layer.

Cross-scale feature correlation

In this sub-module, we explore the correlation of the adjacent feature under various scales. Various parts of the body have distinctive scales. So it is worthwhile to associate different scales of body parts. Same as SFC, we obtain feature blocks S_d^{max} by MaxPooling for each slicing scheme S_d. For this purpose, we design a graph-based self-attention mechanism to aggregate the above information. We construct the graph $G(V, E)$, where V and E represent the set of vertex and edge, respectively. As for vertex set V, $v_{i,j}$ is viewed as the vertex of the graph obtained from slicing scheme $DIVISION$, where i denotes the i-th slicing scheme(S_i) and j denotes the j-th block in S_i. For example, $v_{3,2}$ denotes the second block in the slicing scheme S_3. Besides, we define the weight $Z_{i,j}$ of vertex $v_{i,j}$ as follows:

$$Z_{i,j} = MaxPooling(S^{max}_{i,j}). \tag{6}$$

where $Z_{i,j} \in \mathbb{R}^{c_f}$. As for edge set E, the edge $(v_{i_1,j_1}, v_{i_2,j_2})$ exits if the bounding box of two vertices intersects. In our work, the bounding box(B) is defined as the border of the sliced block. In other words, vertex v_{i_1,j_1} and vertex v_{i_2,j_2} are neighbors when and only when $B_{i1,j1} \bigcap B_{i2,j2} \neq \emptyset$, as shown in Fig. 3. On the basis of the softmax [35], we adopt the graph-based [17] approach to aggregate adjacent features. The attention weights α are obtained by normalizing the aggregation of all neighbors.

$$softmax_{graph-G}(\alpha_{i,j}) = \frac{e^{W_{i,j}}}{\sum_{k \in Nei(V_i)} e^{W_{i,k}}}. \tag{7}$$

where i represents the vertex in graph G for simplicity, $Nei(V_i)$ indicates the set of neighboring vertices for vertex i, and W denotes the weight matrix. In summary, our graph-based self-attention process can be described as follows:

$$GAT(Q, K, V) = softmax_{graph-G}\left(\frac{QK^T}{\sqrt{D}}\right)V. \tag{8}$$

where Query, Key, and Value are obtained from a sequence of vectors through different linear project: $Q = D^{max}W_Q$, $K = D^{max}W_K$, $V = D^{max}W_V$ and $\sqrt{D}$ denotes the regularization terms. Similarly to [36], we also follow a multi-head manner. After GAT, $\hat{Z}_{i,j}$ is the output information of the vertex $v_{i,j}$ aggregated with neighborhoods. As with the SFC sub-modules, we similarly utilize the fully connected layer to extract the cross-scale feature f^d_{cs} from $\hat{Z}_d$, where d denotes the number of block pieces in the slicing scheme S_d.

Loss function

In this paper, we adopt a common loss function for the person Re-ID task, which includes Cross-entropy Loss with label smooth [37] and Hard Triplet Loss [38]. Among these, the cross-entropy loss function can enhance network classification performance.

$$L_{CE}(p, q) = \sum_{i=1}^{k} -p_i \log(q_i). \tag{9}$$

where k, p, and q indicate the number of person categories, predict value, and ground truth, respectively. In addition, the parameter in label smooth is set to 0.1. Hard Triplet Loss can better extract distinguishing identification information, closer to the same identification and farther away from different identification.

$$L_{Tri}(a) = [S^a_p - S^a_n + \alpha]_+. \tag{10}$$

where S^a_p and S^a_n respectively denote positive and negative samples. α is defined as margin distance. Function $[\cdot]_+$ means $max(\cdot, 0)$. Combining Eq. 9 and Eq. 10, the multi-loss function L_{reid} is obtained as follows:

$$L_{reid} = L_{CE} + L_{Tri}. \tag{11}$$

Finally, We leverage features f_{gl}, f^d_{ss} and f^d_{cs} to obtain overall loss function L of whole framework:

$$L = L_{reid}(f_{gl}) + \sum_{d \in DIVISION} (L_{reid}(f^d_{ss}) + L_{reid}(f^d_{cs})). \tag{12}$$

Experiments

In this section, we conduct extensive experiments on commonly public Re-ID datasets with the aim of verifying the effectiveness of our method, which includes Market-1501 [39], CUHK03-NP [40], DukeMTMC [41] and MSMT17 [42]. Firstly, we describe our experimental environment and related evaluation metrics. Then, typical person Re-ID datasets are introduced in the following. Finally, we demonstrate the advantages of our method compared to state-of-the-art models, while validating the effectiveness of each sub-module through ablation experiments.

Implementation details

Training: In our study, ResNet50 is employed as the backbone for end-to-end training, where the stride of the last layer of ResNet50 is set to 1. We set the batch size to 64, in which a total of 16 distinctive identifications, each with 4 different images respectively.

For the image pre-processing step, we standardize the resolution of all images to 256 × 128 pixels to ensure consistency throughout the dataset. Additionally, we employ various augmentation techniques to further enhance the data. These techniques include random cropping, horizontal flipping, and random erasing. By applying these operations, we introduce diversity and variability into the training data, which in turn enhances the model's ability to generalize and improve its overall performance.

For the parameter optimization, we train our model for 180 epochs using the Adam optimizer. Specifically, we employ a warmup strategy where the learning rate is linearly increased from 5×10^{-6} to 4×10^{-4} within the first 10 epochs. Starting from the 50th epoch, the learning rate is adjusted downward every 30 epochs by a factor of 0.4. As for the hyperparameters in the Multi-scale Feature Enhancement (MFE) module, we set R (the size of patch size) as 1 according to the literature [16] and l (the number of layers) as 6. These values are chosen based on empirical observations and experimental results to achieve optimal performance in our model.

Testing: During the testing stage, we extract the feature representation of each image by passing it through our proposed MSAF network. When conducting a query image, we compare the feature representation of the query image with the feature representations of all images in the gallery set using Euclidean Distance as the measurement metric. Based on the similarity scores calculated using the Euclidean Distance, we retrieve the top-K most similar images from the gallery set as the potential matches for the query image. It is important to note that we do not employ re-ranking as a post-processing step in our approach.

Besides, following previous work [16, 43], the cumulative matching characteristic (CMC) and the mean Average Precision (mAP) are adopted in our experiments. The code is implemented on PyTorch with 4 NVIDIA GeForce RTX 2080Ti.

Datasets

We verify the effectiveness of our method by extensive experiments under public Re-ID datasets, including Market-1501 [39], DukeMTMC [41], CUHK03-NP [40], and MSMT17 [42].

Market1501: It is composed of 32,668 images collected on campus from 1,501 pedestrians in total. The dataset contains 12,936 training images of 751 identities, 3,368 query images and 15,913 gallery images of other 750 identities. In this dataset, the pedestrian bounding boxes are detected with the Deformable Part Model (DPM) detector [44].

DukeMTMC: There are a total of 1,402 identities from 8 different cameras in the dataset, which is divided into 702 identities for training and another 702 identities for testing. It contains 16,522 training images, 2,228 query images, and 17,661 gallery images.

CUHK03-NP: In literature [45], the new testing protocol is proposed with 14,097 images of 1,467 pedestrians from 5 pairs of cameras with different views. The 767 identities served as the training set and another 700 identities as the testing set. Besides, the dataset provides two types of bounding boxes generated by manual human annotation and automatic detection with the pedestrian detector, which are simplified as Labeled and Detected, respectively.

MSMT17: In literature [42], a mostly challenging Re-ID dataset with complicated scenes is established. It exploits 15 cameras to obtain images of pedestrians in various weather conditions. Based on Faster RCNN detector [46], it obtains altogether 126,441 images from 4101 identities. Through its random division, it includes 1,041 identities with 32,621 training images and 3,060 identities with 11,659 query images and 82,161 gallery images.

Comparison with state-of-the-art models

We compare our method with the state-of-the-art models and the results are shown in Table 1. We can observe that our model achieves outstanding performance on most datasets.

Market1501: As shown in Table 1, we obtain comparable performance in Market1501 dataset. In detail, our approach achieves 90.3% mAP and 96.2% Rank-1. There is near saturation in the performance of this dataset.

DukeMTMC: As shown in Table 1, it is noticed that our method still achieves the best performance among previous methods on DukeMTMC. In terms of data, our method outperforms NFormer [57] by 0.2% mAP and 2.8% Rank-1. In particular, compared to the previous part-based approaches without correlation under multi-scale adjacent features, e.g., ALignedReID++ [18], GCP-F [31], our method holds a significant advantage.

CUHK03-NP: Table 1 shows that the proposed approach achieves outstanding performance both on Labeled and Detected datasets. In Labeled CUHK03-NP, our approach achieves 85.5% mAP and 83.4% Rank-1. Meanwhile, we obtain 80.7% mAP and 83.3% Rank-1 on Detected CUHK03-NP. More specifically, when comparing our proposed method to advanced HAT [16] and GCP-F [31], we observe significant improvements. For instance, on the Labeled CUHK03-NP dataset, our method achieves the improvement of 3.4% mAP and 2.9% Rank-1. Similarly, on the Detected CUHK03-NP dataset, we observe the improvement of 5.2% mAP and 4.2% Rank-1. These results demonstrate the superior performance of our approach compared.

MSMT17: As shown in Table 1, our approach obtains 63.4% mAP and 83.7% Rank-1 on MSMT17. Specifically, in terms of the Rank-1 metric, our method surpasses TransReID [32] by a margin of 1.0%. However, our mAP result falls slightly short by 0.5% compared to TransReID. When compared to Nformer [57], we outperform them by at least 3.6% and 6.4% in Rank-1 accuracy and mAP, respectively. Therefore, our proposed approach is still competitive.

The MSMT17 dataset is notable for its collection of 12 outdoor cameras and 3 indoor cameras, resulting in

Table 1 Comparison with state-of-the-art methods on Market1501, DukeMTMC, CUHK03-NP, MSMT17 datasets. **Red**, **green**, and **blue** represent the top-3 ranking scores respectively

			CUHK03-NP							
Methods	Market-1501		Labeled		Detected		DukeMTMC		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
ABDNet [47]	88.3	95.6	-	-	-	-	78.6	89.0	60.8	82.3
SFT [48]	87.5	94.1	-	-	71.7	74.3	79.6	90.0	58.3	79.0
HPM [49]	82.7	94.2	-	-	57.5	63.9	74.3	86.6	-	-
Pyramid [24]	88.2	95.7	76.9	78.9	74.8	78.9	79.0	89.0	-	-
CAMA [50]	84.5	94.7	66.5	70.1	64.2	66.6	72.9	85.8	-	-
JDGL [51]	86.0	94.8	-	-	-	-	74.8	86.6	52.3	77.2
AlignedReID++ [18]	79.1	91.8	-	-	59.6	61.5	69.7	82.1	43.7	69.8
HBFPNet [22]	**89.8**	**95.8**	**79.4**	**81.3**	**77.5**	**80.0**	80.2	89.5	-	-
ISP [52]	88.6	95.3	74.1	76.5	71.4	75.2	80.0	89.6	-	-
RGA-SC [29]	88.4	**96.1**	77.4	81.1	74.5	**79.6**	-	-	57.5	80.3
SNR [53]	84.7	94.4	-	-	-	-	73.0	85.9	-	-
GCP-F [31]	88.9	95.2	75.6	77.9	69.6	74.4	78.6	89.7	-	-
HAT [16]	89.5	95.6	**80.0**	**82.6**	**75.5**	79.1	**81.4**	**90.4**	61.2	82.3
CDNet [54]	86.0	95.1	-	-	-	-	76.8	88.6	54.7	78.9
PAT [55]	88.0	95.4	-	-	-	-	78.2	88.8	-	-
APD [56]	89.1	**95.8**	77.2	79.9	75.3	78.1	81.1	**90.7**	61.2	82.4
3DTANet [1]	89.6	95.3	75.2	80.2	68.9	75.2	78.4	89.9	46.7	76.6
TransReID [32]	88.0	94.7	-	-	-	-	81.2	90.1	**63.9**	**82.7**
AAformer [14]	88.0	95.4	79.0	80.3	77.2	78.1	80.9	90.1	**65.6**	**84.4**
NFormer [57]	**91.1**	94.7	78.0	77.2	74.7	77.3	**83.5**	89.4	59.8	77.3
Ours	**90.3**	**96.2**	**83.4**	**85.5**	**80.7**	**83.3**	**83.7**	**92.9**	**63.4**	**83.7**

a wide variation in the style of pedestrian images across different cameras. Consequently, the incorporation of camera/view variations plays a crucial role in capturing non-visual information. However, different from TransReID[32] and AAformer [14], our method disregards the consideration of camera/view variations, which may limit its performance in scenarios where such variations are significant. In our work, despite the lack of variations as prior knowledge, robust identification information is maintained due to discriminative multi-scale adjacent feature extraction. As a result, our approach remains competitive in the MSMT17 dataset.

Ablation studies

For the purpose of demonstrating the effectiveness of the proposed components, ablation studies are performed on the DukeMTMC dataset. We adopt pure ResNet50 as the baseline and set the stride of the last layer to 1. Besides, we incorporate Cross-entropy Loss and Hard Triplet Loss to train the network in all ablation experiments. To fully illustrate the capacity of the components to represent features under different granularities, we take $DIVISION = \{S_1, S_3, S_5, S_7\}$ in some experiments.

Effect of MFE and JPFA in MSAF: As shown in Table 2, we can observe that the MFE improves by 7.4%/5.6% on mAP/Rank-1 compared to baseline. It is sufficient to demonstrate that a multi-feature extraction approach can deliver effective performance improvements. Table 2 shows that the JPFA outperforms baseline with 2.5% in mAP and 3.4% in Rank-1. From the above experimental result, we conclude that our proposed JPFA enables the effective extraction of local features. This adequately describes the benefits of multi-scale adjacent feature aggregation for mining identification information.

Effect of intra-scale and inter-scale in MFE: As shown in Table 3, we further analyze the effect of each feature extraction scheme: intra-scale and inter-scale.

For intra-scale, the results show that the intra-scale feature extraction method can enhance the performance by 1.8% and 1.6% in mAP and Rank-1, respectively. And it also indicates that the intra-scale scheme significantly improves the mAP metric from 76.7% to 83.0% and the Rank-1 metric from 88.5% to 91.2%. Based on the above results, we infer that learning more contextual information within the same level can be extremely effective, which allows for better inter-scale feature extraction. For inter-scale, it shows that the inter-scale scheme gains 1.1% in mAP and 1.9% in Rank-1 benefit compared to pure baseline. Besides, better performance is obtained by joining inter-scale with intra-scale. In detail, the MFE achieves a large margin(+5.6 % in mAP and 3.0% in Rank-1) over a pure intra-scale scheme. Furthermore, we compare the performance of MFE and state-of-the-art models shown in Table 1. The comparison shows that it outperforms the second-best model 0.5% in mAP and 1.6% in Rank-1 on the DukeMTMC dataset. Through the above analysis, it is strongly demonstrated that integrating low-level detail information and high-level semantic information is able to extract pedestrian features effectively. Also, it illustrates the ability of MFE to obtain discriminative-specific information.

Table 2 Ablation studies of our modules on DukeMTMC. The slicing scheme is set as $DIVISION = \{S_1, S_3, S_5, S_7\}$

Method	mAP	Rank-1	Rank-5	Rank-10
Baseline(B)	75.6	86.6	94.3	96.4
B+MFE	83.0	91.2	96.0	**97.4**
B+JPFA	78.1	90.0	95.1	96.6
Ours(B+MFE+JPFA)	**83.6**	**92.1**	**96.3**	97.2

Table 3 Ablation studies of MFE on DukeMTMC. The slicing scheme is set as $DIVISION = \{S_1, S_3, S_5, S_7\}$

Method	mAP	Rank-1	Rank-5	Rank-10
Baseline(B)	75.6	86.6	94.3	96.4
B+MFE(Intra)	77.4	88.2	94.6	96.5
B+MFE(Inter)	76.7	88.5	94.4	96.0
B+MFE(Intra+Inter)	**83.0**	**91.2**	**96.0**	**97.4**

Table 4 Ablation studies of JPFA on DukeMTMC. The slicing scheme is set as $DIVISION = \{S_1, S_3, S_5, S_7\}$

Method	mAP	Rank-1	Rank-5	Rank-10
Baseline	75.6	86.6	94.3	96.4
w/o JPFA	83.0	91.2	96.0	97.4
w/o CFC	82.6	91.5	**96.4**	**97.5**
w/o SFC	83.1	91.6	96.1	97.0
Ours	**83.6**	**92.1**	96.3	97.2

Effect of CFC and SFC in JPFA: We also conduct experiments to evaluate the benefit of JPFA, which contains SFC and CFC sub-modules.

As shown in Table 4, we separately verify the effectiveness of SFC and CFC. For SFC, the accuracy decreases by 0.5% mAP and 0.5% Rank-1 without the SFC sub-module. It suggests that fusing adjacent features of the same slicing scheme is critical. Meanwhile, the CFC sub-module which is based on cross-scale adjacent features can improve the performance of the whole framework by 1.0% mAP and 0.6% Rank-1. These results show that cross-scale feature extraction adapts effectively to the proportions of different body parts. The above experiments have shown that JPFA effectively compensates for semantic information ambiguity and background interference due to intensive partition.

Experiment analysis

Analysis of feature extraction in MFE from different-level: As shown in Table 5, we further explore the performance of feature extraction from different-level. Firstly, we denote the l-th layer of ResNet50 as X_l. We conduct some experiments of different combinations with $\{X_3, X_4, X_5\}$. Compared $\{X_3, X_4\}$ to $\{X_3, X_4, X_5\}$, it can be observed that the performance of mAP and Rank-1 significantly decreases 5.7% and 3.1% respectively without high-level feature. Through additional insights, $\{X_3, X_4, X_5\}$ combination is the worst performance of all the two different levels selected. It adequately indicates that high-level semantic information plays a fundamental role in identification verification. Similarly, compared $\{X_4, X_5\}$ to $\{X_3, X_4, X_5\}$, it shows that low-level feature can enhance the accuracy by 1.5% mAP and 0.9 % Rank-1. From the above comparison, we infer that the inclusion of only higher-level semantic information does not allow for the extraction of rich identification information. In other words, it demonstrates the importance of low-level detail information to some extent. The comparison of $\{X_3, X_5\}$ and $\{X_3, X_4, X_5\}$ illustrates that our model performs better with the assistance of mid-level features. Therefore, we have experimentally shown that the information is required at each level. It also shows that the integration of detailed and semantic information can contribute to improved discrimination and robustness.

Table 5 Ablation studies of feature aggregation layer in MFE on DukeMTMC. The digit represents the layer of ResNet50. The slicing scheme is set as $DIVISION = \{S_1, S_3, S_5, S_7\}$

Method	mAP	Rank-1	Rank-5	Rank-10
{3, 4}	77.9	89.0	95.0	96.4
{3, 5}	81.6	91.2	95.9	**97.2**
{4, 5}	82.1	91.2	96.0	**97.2**
{3, 4, 5}	**83.6**	**92.1**	**96.3**	**97.2**

Table 6 Ablation studies of the slicing scheme in JPFA on DukeMTMC

Method	mAP	Rank-1	Rank-5	Rank-10
$\{\emptyset\}$	83.0	91.2	96.0	97.4
$\{S_1, S_3\}$	83.0	91.8	96.1	97.1
$\{S_1, S_4\}$	82.7	91.8	96.0	97.2
$\{S_1, S_5\}$	83.0	92.0	96.1	97.4
$\{S_1, S_7\}$	82.8	91.6	96.5	97.2
$\{S_1, S_3, S_4\}$	83.1	92.2	96.4	97.4
$\{S_1, S_3, S_5\}$	83.3	92.4	96.1	97.3
$\{S_1, S_4, S_5\}$	83.4	92.2	96.5	97.6
$\{S_1, S_3, S_5, S_7\}$	83.6	92.1	96.3	97.2
$\{S_1, S_3, S_4, S_5\}$	**83.7**	**92.9**	**96.5**	**97.7**

The digit represents the number of blocks uniformly sliced from the feature map. Especially, Ø means no slicing scheme

Analysis of slicing scheme in JPFA: As shown in Table 6, we further compare the effectiveness of various combinations with different slicing schemes. Following previous part-based methods, we uniformly slice the feature map into horizontal blocks of equal size. In particular, the empty set

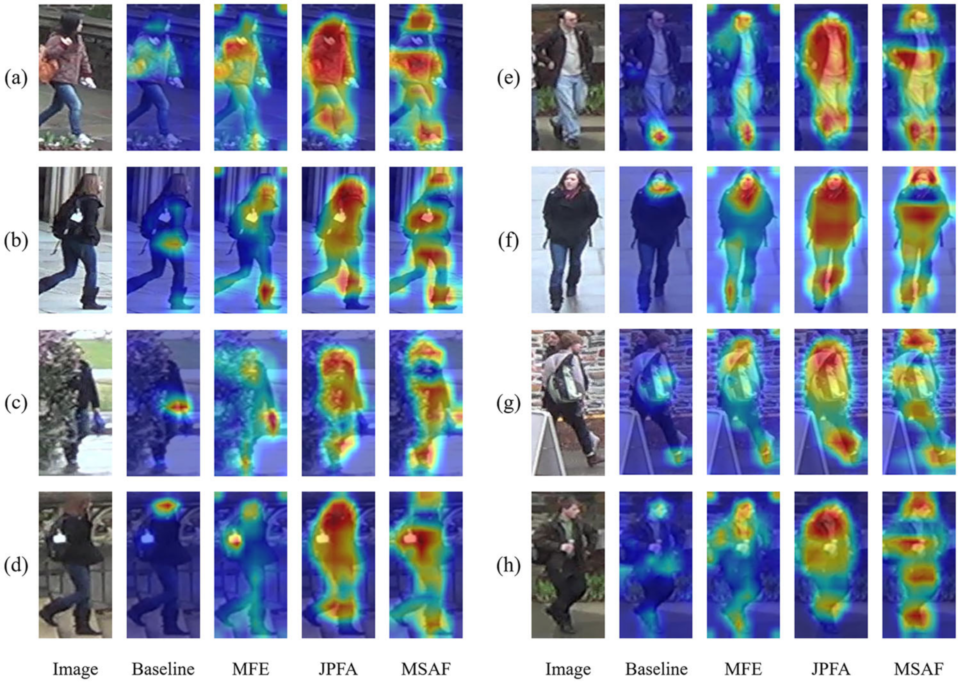

Fig. 4 Visualization of our proposed method. From left to right, each column donates input image, the visualization of the baseline, MFE, JPFA, MSAF, respectively

indicates that our slicing strategy is not employed and only the global feature is utilized. For the metric of Rank-1, it is seen that the results of $\{S_1, S_3\}$ and $\{S_1, S_5\}$ outperform no slicing strategy by at least 0.6% margin. As illustrated by extensive experiments, the part-based approach has the capability to extract fine-grained local features with visible effects. However, excessively fine-grained local features prevent the model from extracting discriminative features and lead to network redundancy. From the comparison of $\{S_1, S_5\}$ and $\{S_1, S_7\}$, we observe reductions in performance of mAP and Rank-1 by 0.2% and 0.4% respectively. On the other hand, the performance of $\{S_1, S_3, S_4, S_5\}$ surpasses $\{S_1, S_3, S_5, S_7\}$ by 0.1% mAP and 0.8% Rank-1. Therefore, as the number of slices increases, we infer that the lack of semantic information within each block leads to slight or even deteriorating improvements in accuracy. Furthermore, we clearly notice that the addition of various rational slicing options can enhance the performance, e.g., $\{S_1, S_3\}$, $\{S_1, S_3, S_4\}$ and $\{S_1, S_3, S_4, S_5\}$. Through comparison of the above experimental results, it indicates that incorporating diverse partition schemes effectively adapts to the proportion of different limbs, while appropriately alleviating semantic information ambiguity and background interference caused by intensive partitions.

Visualization

As shown in Fig. 4, we visually compare the feature maps of the proposed method with CAM [58] visualization. For each identity image, it presents the result of the input image and the visualization of the baseline, MFE, JPFA, and MSAF. It can be seen that our MFE module provides access to more pedestrian detail than the baseline. In addition, as viewed in Fig. 4a, c, MFE can effectively filter the occlusion object to extract more specific features, e.g., handbag. From the column of JPFA, it is clear that the whole body of the feature map has been focused on. It demonstrates that the interactions with diverse scales effectively address semantic information ambiguity and background interference caused by intensive partitions. And it adapts to the different proportions of body parts. However, feature aggregation is based on feature extraction. It leads to the lack of salient clues at the low-level by applying a partition scheme in the high-level and abstract representation space. For example, Fig. 4c illustrates that JPFA is concerned about a small amount of background noise, and Fig. 4d shows limited attention to key body information. From the last column, the result of visualization shows that the combination of MFE and JPFA components provides the ability to extract specific discriminative information with different structural proportions. Therefore, our method achieves better performance than many state-of-the-art methods.

Conclusion

In this paper, we proposed a novel learning Discriminative Multi-scale Adjacent Feature (MSAF) model for person Re-ID to effectively alleviate semantic information ambiguity and background interference. It contained feature extraction and feature aggregation. In the feature extraction stage, the MFE module was designed by combining CNN and Transformer to effectively obtain discriminative specific information from different-layer. In the feature aggregation stage, the JPFA mechanism was revealed to implement adjacent feature aggregation with diverse scales by fusing the SFC and CFC sub-modules. Besides, it could further extract specific discriminative information from different structural proportions to enhance the robustness of identification representation. Ablation experiments proved the validity of each component of the model. Experimental analysis and visualization illustrated the rationality and interpretability of our scheme. The extensive experiments on four public Re-ID datasets demonstrated that our method outperformed many state-of-the-art models. Further work will be carried out to find higher efficiency as well as simplify the complexity of the network for practical.

Acknowledgements This work is partially supported by the National Natural Science Foundation of China under Grant (No.61906168, No.62272267), Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020023, LZ23F020001), Joint Funds of the Zhejiang Provincial Natural Science of China under Grant (No.LZJWZ22E090001), the Construction of Hubei Provincial Key Laboratory for Intelligent Visual Monitoring of Hydropower Projects (Grant No. 2022SDSJ01), the Hangzhou AI major scientific and technological innovation project (Grant No. 2022AIZD0061), Yongjiang Talent Introduction Programme(2022A-234-G) and Quzhou Science and Technology Projects(2022K91).

Author Contributions Conceptualization: MQ and SC Data Curation: MQ, FH and SC Investigation: YY, XZ and SC Software and validation: MQ, SC, FH and XZ Visualization: SC, YY and XZ writing—original draft preparation: MQ and SC writing—review and editing: FH, YY and XZ Formal analysis: MQ, FH and SC Funding acquisition: SC, YY and XZ Supervision: SC and FH Project administration: SC, YY and XZ All authors have read and agreed to the published version of the manuscript.

Data availability Partial or complete data, models, or code generated during the research process can be obtained from the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflict of interests.

References

1. Huang Y, Lian S, Zhang S, Hu H, Chen D, Su T (2020) Three-dimension transmissible attention network for person re-identification. IEEE Trans Circ Syst Video Technol 30(12):4540–4553
2. Chan S, Du F, Tang T, Zhang G, Jiang X, Guan Q (2023) Parameter sharing and multi-granularity feature learning for cross-modality person re-identification. Complex & Intelligent Systems, 1–14
3. Chan S, Liu Y, Pan X, Lei Y (2023) Person re-identification based on feature fusion in ai system. International Journal of Humanoid Robotics, 2350004
4. Wang G, Yuan Y, Chen X, Li J, Zhou X (2018) Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282
5. Yang Q, Yu H-X, Wu A, Zheng WS (2019) Patch-based discriminative feature learning for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3633–3642
6. Qi M, Chan S, Hang C, Zhang G, Li Z (2023) Fine-grained learning for visible-infrared person re-identification. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 2417–2422. IEEE
7. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 480–496
8. Zhang X, Luo H, Fan X, Xiang W, Sun Y, Xiao Q, Jiang W, Zhang C, Sun J (2017) Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184
9. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inform Process Syst **25**
10. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I(2017) Attention is all you need. Adv Neural Inform Process Syst **30**
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
13. El-Nouby A, Neverova N, Laptev I, Jégou H (2021) Training vision transformers for image retrieval. arXiv preprint arXiv:2102.05644
14. Zhu K, Guo H, Zhang S, Wang Y, Huang G, Qiao H, Liu J, Wang J, Tang M (2021) Aaformer: auto-aligned transformer for person re-identification. arXiv preprint arXiv:2104.00921
15. Peng Z, Huang W, Gu S, Xie L, Wang Y, Jiao J, Ye Q (2021) Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 367–376
16. Zhang G, Zhang P, Qi J, Lu H (2021) Hat: Hierarchical aggregation transformers for person re-identification. In: Proceedings of the 29th ACM international conference on multimedia, pp. 516–525
17. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv:1710.10903
18. Luo H, Jiang W, Zhang X, Fan X, Qian J, Zhang C (2019) Alignedreid++: dynamically matching local information for person re-identification. Pattern Recognit 94:53–61
19. Yi D, Lei Z, Liao S, Li SZ (2014) Deep metric learning for person re-identification. In: 2014 22nd international conference on pattern recognition, pp. 34–39. IEEE

20. Yu Q, Chang X, Song Y-Z, Xiang T, Hospedales TM (2017) The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. arXiv preprint arXiv:1711.08106
21. Chen X, Fu C, Zhao Y, Zheng F, Song J, Ji R, Yang Y (2020) Salience-guided cascaded suppression network for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3300–3310
22. Liu Z, Zhang L, Yang Y (2020) Hierarchical bi-directional feature perception network for person re-identification. In: Proceedings of the 28th ACM international conference on multimedia, pp. 4289–4298
23. Martinel N, Foresti GL, Micheloni C (2020) Deep pyramidal pooling with attention for person re-identification. IEEE Trans Image Process 29:7306–7316
24. Zheng F, Deng C, Sun X, Jiang X, Guo X, Yu Z, Huang F, Ji R (2019) Pyramidal person re-identification via multi-loss dynamic training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8514–8522
25. Xia BN, Gong Y, Zhang Y, Poellabauer C (2019) Second-order non-local attention networks for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3760–3769
26. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (Csur) 40(2):1–60
27. Cui J, Chan S, Mu P, Tang T, Zhou X (2023) Pure detail feature extraction network for visible-infrared re-identification. Intelligent Automation & Soft Computing **37**(2)
28. Chan S, Cui J, Wu Y, Wang H, Bai C (2023) Visible-xray cross-modality package re-identification. In: 2023 IEEE international conference on multimedia and expo (ICME), pp. 2579–2584. IEEE
29. Zhang Z, Lan C, Zeng W, Jin X, Chen Z (2020) Relation-aware global attention for person re-identification. In: Proceedings of the Ieee/cvf conference on computer vision and pattern recognition, pp. 3186–3195
30. Hou R, Ma B, Chang H, Gu X, Shan S, Chen X (2019) Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9317–9326
31. Park H, Ham B (2020) Relation network for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11839–11847
32. He S, Luo H, Wang P, Wang F, Li H, Jiang W (2021) Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 15013–15022
33. Ng T, Balntas V, Tian Y, Mikolajczyk K (2020) Solar: second-order loss and attention for image retrieval. In: European conference on computer Vision, pp. 253–270. Springer
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778
35. Sun Y, Wang X, Tang X (2014) Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1891–1898
36. Chen C-FR, Fan Q, Panda R (2021) Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International conference on computer vision, pp. 357–366
37. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826
38. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823
39. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp. 1116–1124
40. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 152–159
41. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE international conference on computer vision, pp. 3754–3762
42. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 79–88
43. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC (2021) Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence
44. Felzenszwalb P, McAllester D, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE conference on computer vision and pattern recognition, pp. 1–8 . Ieee
45. Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1318–1327
46. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**
47. Chen T, Ding S, Xie J, Yuan Y, Chen W, Yang Y, Ren Z, Wang Z (2019) Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 8351–8361
48. Luo C, Chen Y, Wang N, Zhang Z (2019) Spectral feature transformation for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4976–4985
49. Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z, Huang T (2019) Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 8295–8302
50. Yang W, Huang H, Zhang Z, Chen X, Huang K, Zhang S (2019) Towards rich feature discovery with class activation maps augmentation for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1389–1398
51. Zheng Z, Yang X, Yu Z, Zheng L, Yang Y, Kautz J (2019) Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2138–2147
52. Zhu K, Guo H, Liu Z, Tang M, Wang J (2020) Identity-guided human semantic parsing for person re-identification. In: European conference on computer vision, pp. 346–363. Springer
53. Jin X, Lan C, Zeng W, Chen Z, Zhang L (2020) Style normalization and restitution for generalizable person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3143–3152
54. Li H, Wu G, Zheng W-S (2021) Combined depth space based architecture search for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6729–6738
55. Li Y, He J, Zhang T, Liu X, Zhang Y, Wu F (2021) Diverse part discovery: Occluded person re-identification with part-aware trans-

former. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2898–2907

56. Lai S, Chai Z, Wei X (2021) Transformer meets part model: Adaptive part division for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4150–4157
57. Wang H, Shen J, Liu Y, Gao Y, Gavves E (2022) Nformer: Robust person re-identification with neighbor transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7297–7307
58. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929