Pattern Recognition 150 (2024) 110362

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Multi-granularity Cross Transformer Network for person re-identification

Yanping Li [a,b], Duoqian Miao [a,b,*], Hongyun Zhang [a,b], Jie Zhou [c], Cairong Zhao [a,b]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[b] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China
[c] College of Computer Science and Software Engineering, Shenzhen University, Nanshan District, Shenzhen City, Guangdong Province, China

ARTICLE INFO

Keywords:
Person re-identification
Cross transformer
Multi-granularity

ABSTRACT

Person re-identification (Re-ID) aims to retrieve the same person in the gallery. Transformers have been introduced to the Re-ID task due to their excellent ability to model long-range dependency. However, due to the properties of the global attention mechanism, they are less effective in capturing the discriminative local semantics of pedestrians compared to convolutional operations. To address this issue, we present a Multi-granularity Cross Transformer Network (MCTN) that progressively learns salient features of different local structures in a global context. Specifically, we first utilize a Multi-granularity Convolutional Layer (MCL) to investigate salient pedestrian features at various granularities. On this basis, we propose a Pyramidal Cross Transformer learning layer (PCT), which contains a pyramidal division of pedestrian image feature maps, differentiated feature extraction of different parts of pedestrians, and cross attention to exploring the local–global relationship of the feature map. It allows effective mining of local information in the global structure from a coarse-to-fine perspective. Furthermore, to enhance the interaction between low-level detailed features and high-level semantic features, a Hierarchical Aggregation Strategy (HAS) is introduced to fuse features learned by cross attention learning at different stages. Pedestrian features learned in shallow layers will serve as global priors for semantics learning in deep layers. We evaluate our method on four large-scale Re-ID datasets, and the experimental results reveal that the proposed method outperforms the state-of-the-art methods.

1. Introduction

Person Re-identification (Re-ID) plays a crucial role in modern intelligent surveillance techniques, such as pedestrian retrieval and behavior analysis [1–4]. However, Re-ID suffers from various challenges such as occlusion, low resolution, and viewpoint/pose/domain/clothing/illumination changes. Thus, it has drawn the attention of many academics.

To learn the discriminative features of pedestrian images, researchers attempt to design effective structures that are robust to the aforementioned challenges. The methods can be roughly cast into three categories. CNN-based methods, pure Transformer-based methods, and CNN+Transformer-based methods. In earlier times, some researchers [5–7] proposed to use the successful architecture of the Convolutional Neural Network (CNN) in other computer vision tasks for extracting robust features of pedestrian images. IDE [5] proposed a convolutional siamese network to predict the IDs of two given pedestrian images and their similarity score, which is composed of the popular pre-trained CNN models. Mudeep [6] devised a multi-scale attention to obtain the attention maps at each scale. RGA [7] proposed relation-aware global attention for better attention learning. Despite the fact that they have yielded promising results in some specific circumstances, they are not robust enough due to limited receptive fields of CNN, single-granularity global features learning, and fewer interactions between detail information in shallow layers and semantic information in deep layers.

With the significant successes achieved by Transformers in many visual tasks, they have also been incorporated into the person Re-ID field. He et al. [8] proposed a pure Transformer architecture to obtain diversified features by rearranging the patch embeddings and mitigating feature bias towards camera variants. Ye et al. [9] proposed a self-supervised method with a channel-wise Transformer to alleviate the domain gap between models pre-trained on ImageNet and ReID datasets. Zhang et al. [10] analyzed the limitations of ViTs in capturing high-frequency components of pedestrian images, and proposed to enhance high-frequency components and drop low-frequency components. Compared with CNN-based approaches, Transformer-based methods excel at learning comprehensive global features of pedestrians. Nevertheless, they may fall short in effective local semantics learning of pedestrians, which can limit the overall performance of these networks.

[*] Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai 201804, China.
E-mail address: dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.patcog.2024.110362
Received 5 December 2023; Received in revised form 16 January 2024; Accepted 19 February 2024
Available online 21 February 2024

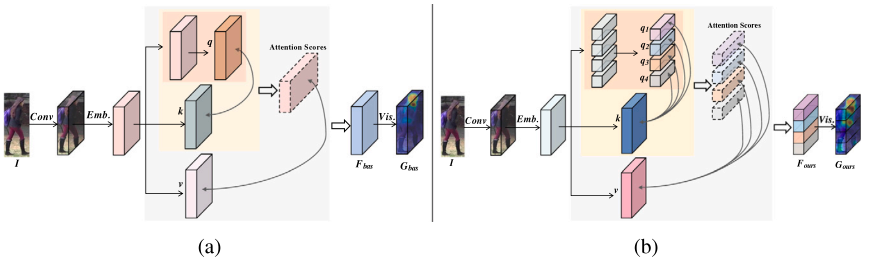


Fig. 1. Comparison between the transformer structure in the baseline model and the proposed MCTN. "$Conv$" denotes convolutional operation; "$Emb.$" is the input embedding; "$Vis.$" is visualization using smooth grad-cam++. For simplicity, we omit the $Norm$ operation and residual connection of the Transformer structure. (a) The process of the baseline model. (b) The process of the proposed MCTN.

More recently, the integration of CNN and Transformer for person ReID has become a current research hotspot. Wang et al. [11] proposed the Neighbor Transformer Network (Nformer), employing a CNN for feature extraction across all images. The landmark agent attention module and reciprocal neighbor softmax function module model interactions among all images, reducing irrelevant features and enhancing overall robustness. Xu et al. [12] used CNN to extract the pose information of pedestrians and built local features. A directed graph is constructed by semantic features of the same part in image pairs and a feature recovery transformer is developed to restore the occluded features of the pedestrians. Liu et al. [13] devised a dual-path structure to extract pedestrian features from CNN and Transformer. The spatial and temporal information in videos is further captured through a complementary content attention module and hierarchical temporal aggregation module. Zhang et al. [14] developed the Hierarchical Aggregation Transformer (HAT) method, where the TFC is inserted into each stage of the CNN network to integrate the output features from the TFC in the previous scale and the current stage of CNN, generating global priors for next scales. These methods can simultaneously exploit the strengths of both CNN and Transformer to learn discriminative features of pedestrian images.

Technically, our method belongs to the kind of CNN+Transformer. Compared with other approaches, our method relies on the concept of multi-granularity feature learning, effectively integrating the multi-granularity features learned within CNN with Transformers, ultimately yielding discriminative global and local features for pedestrians. Building on the baseline model HAT [14], we have introduced several innovative and key design concepts. The Multi-granularity Convolutional Layer (MCL) is utilized to investigate salient pedestrian features at various granularities, which serves as the prerequisite of subsequent processes to mine local information in the global structure from a coarse-to-fine perspective. To enhance the Transformer's awareness of local context, we propose a Pyramidal Cross Transformer learning layer (PCT) for mining local information within the global structure, which includes the coarse-to-fine division of pedestrian image feature maps, differentiated feature extraction of different parts of pedestrians, and cross attention to exploring the local–global relationship of the feature maps. To show the effectiveness of the proposed MCTN, we take an example in Fig. 1. As shown in Fig. 1(a), the Transformer structure of the baseline model [14] directly takes the global features as input, which can only produce the most salient pedestrian features F_{bas}, i.e., the heat of the attention map G_{bas} is located in only a limited portion of the pedestrian image. In contrast, in Fig. 1(b), the proposed MCTN takes the global features as input, followed by a uniform division of global features to learn discriminative semantic information implied in various local structures of the pedestrian image, producing more comprehensive and robust pedestrian features F_{ours}, i.e., the heat of the attention map G_{ours} is distributed across the entire pedestrian images.

Our contributions are summarized as follows:

(1) We propose a Multi-granularity Cross Transformer Network (MCTN) to solve the problem that the existing methods ignore the discriminative semantic information implied in various local structures in the global feature maps.

(2) To take full advantage of the Transformer's capabilities in the person Re-ID task, we propose a novel dedicated module, i.e., the Pyramidal Cross Transformer learning layer (PCT), which enables the network to obtain rich and diverse clues for Re-ID. Meanwhile, we propose a Hierarchical Aggregation Strategy (HAS) to enhance the fusion of low-level details and high-level semantics.

(3) MCTN achieves the state-of-the-art performance on popular Re-ID datasets, i.e., Market-1501, DukeMTMC-reID, CUHK03 and MSMT17. Additionally, conducting extensive ablation studies will provide valuable insights into the design of the network for future advancements.

2. Related work

To address the challenges mentioned above, a number of methods have been proposed, which can be divided into three categories [15], i.e., global feature-based representation learning, local feature-based representation learning and auxiliary feature-based representation learning. In this section, we first briefly overview these methods and then introduce the Transformer and Multi-granularity-related methods in Re-ID.

2.1. Global feature-based representation learning

Global feature-based representation learning aims to generate comprehensive and discriminative features for each individual pedestrian image. The global features are usually obtained by imposing Global Max Pooling (GMP) or Global Average Pooling (GAP) on the feature maps learned by the convolutional neural networks. IDE [5] is one of the early representative studies that attempted to view the Re-ID problem as a classification and verification task, which simultaneously learns discriminative embeddings for pedestrian images and a similarity metric. Effective IDE-based improvements include Label Smoothing [16], SphereReID [17], etc., which are widely utilized in many current mainstream approaches. Luo et al. [18] proposed a powerful baseline that includes a bag of tricks to enhance discriminative global feature learning. Recently, Ye et al. [15] designed a new baseline termed AGW, which consists of Non-local Attention (Att) block, Generalized-mean (GeM) pooling, and Weighted Regularization Triplet (WRT) loss. Qian et al. [19] developed a leader-based multi-scale attention deep architecture that dynamically determines the importance of the discriminative features extracted at each scale. Zhang et al. [7] presented a Relation-aware Global Attention (RGA) to grasp information of global scope. MEMF [20] incorporates multi-level-attention blocks into a multi-layer-feature fusion architecture, enabling the extraction of representative and rich features. SDN [21] proposes a global-correlation and a local-correlation attention to capture inter-image and intra-image dependencies, respectively. DAAF-BoT [22] utilizes a holistic attention branch for global awareness, focusing on persons to reduce background influence, and a partial attention branch for local awareness, decoupling features into groups responsible for different body parts. Zhang et al. [23] proposed a Heterogeneous Convolutional Network (HCN) to jointly learn the appearance information of pedestrian images and their correlations. Even though CNN-based backbone networks and some effective modules are capable of extracting robust global feature representations and achieving satisfactory matching performance in some specific circumstances, global features cannot successfully cope with complex scenes, such as occlusions.

2.2. Local feature-based representation learning

In order to compensate for the shortcomings of global features in, e.g., occlusion scenarios, a number of methods based on pose estimation, background segmentation, horizontal division, etc., have been proposed. Zhao et al. [24] proposed the Spindle Net, the first attempt to introduce the human body structure information into the Re-ID framework. M. Kalayeh et al. [25] developed SPReID that used human semantic parsing to explore local visual cues for Re-ID. This type of method makes the network relatively more complex as it requires the introduction of other sub-networks (e.g., OpenPose [26] for pose estimation) for human parsing. In contrast, the horizontal division is much more flexible. Sun et al. [27] proposed to divide the feature maps into several stripes before conducting the global max pooling to generate local features. This design has proven to be remarkably effective and has become a standard for many local feature-based methods. Wang et al. [28] developed a multi-branch strategy with one branch for global feature learning and two branches for local feature learning. The multiple granularities in this paper refer to the division of body parts with different sizes. To capture the relationship among the different body parts, the second-order non-local attention [29] is designed to enhance feature learning. Although this type of method is more flexible, it is sensitive to heavy occlusions as well.

2.3. Auxiliary feature-based representation learning

The auxiliary features used in Re-ID mainly involve semantic attributes, viewpoint information, domain information, and data augmentation [15]. Lin et al. [30] first incorporated the semantic attributes learning into the Re-ID networks. Chang et al. [31] proposed a Multi-Level Factorization Net (MLFN) that factorizes the visual appearance of a person into multiple semantic feature learning. Lin et al. [32] exploited the camera consistency-aware to learn both feature representation and image matching simultaneously. The data augmentation mainly includes a random erasing strategy [33] and using the images generated by GAN for training [34,35]. The former aims to mimic the occlusion that often occurs in the real world. The latter is designed to generate pedestrian images with a variety of challenges, e.g., occlusions, viewpoint/pose/domain/clothing/illumination changes, in order to render the network robust to these disruptions for Re-ID. Those methods enrich the data types of the training images, allowing the network to be more generalizable when tested.

2.4. Transformer and multi-granularity in Re-ID

Transformer is an advanced design that originated in Natural Language Processing (NLP) [36] and is now widely used in various areas of computer vision with promising results. The Vision Transformer (ViT) [37] and its variants [38–40] divided the image into patches, which can be considered as the word tokens in NLP. For the Re-ID task, He et al. [8] proposed a pure transformer framework for person Re-ID, which contains two novel modules, i.e., jigsaw patch module (JPM) and side information embeddings (SIE). The philosophy of these two designs stems from the auxiliary feature-based methods. Recently, many researchers have considered combining the respective strengths of CNN and Transformer for application in Person ReID. Chen et al. [41] designed an attended structure representation based on CNN for learning structure-related features of pedestrian appearance and developed transformer-based part interaction for exploring the contextual and structural relationships between part levels using a node-level partitioning strategy. Wang et al. [11] proposed the NFormer, where the CNN is used to extract features from all pedestrians. The landmark agent attention module and the reciprocal neighbor softmax function module are designed based on the Transformer to yield robust feature representations. Similarly, Lai et al. [42] also utilized the CNN to extract pedestrian features. Meanwhile, a Transformer-based part merge module and a part mask generation module are proposed to capture the corresponding areas of two different samples and exhibit robustness to scale and shift variations. Zhang et al. [14] introduced a Transformer-based Feature Calibration (TFC) module at each stage of the CNN network to fuse the output features of the backbone network at each stage with those of the previous stage TFC module. Those methods have achieved promising results in the Re-ID task due to the long-range dependencies modeling ability of the attention mechanism.

Multi-granularity feature learning has also been shown to play an important role in person Re-ID. MGN [28] and MGCA [43] both considered using differentiated division for feature maps from different branches to obtain the local features with diverse granularities, without any operation for further feature learning. Chen et al. [44] proposed a Saliency and Granularity Mining Network (SGMN) for solving video-based person ReID. The “granularity” mainly refers to mining small granularity information of pedestrians in each frame such as logo, and shoes, using the temporal channel-relation module. Zhang et al. [45] proposed to enable the gradual expansion of the size of feature maps to obtain multiple resolution features, followed by a multi-pool feature extractor for producing more discriminative features by fusing high and low-level features. Gong et al. [46] designed the network as a multi-stream structure. By constructing a global feature stream and a part feature stream, coarse-grained global information and fine-grained localized body part information are preserved, respectively. Zhang et al. [47] divided the feature maps into groups along the channel dimension. Each group represents a granularity and a spatial average pooling operation is performed for each granularity at different scales. Wang et al. [48] proposed a Receptive Multi-Granularity Learning (RMGL) method, employing a multi-branch network and incorporating an activation balanced pooling strategy to achieve adaptive partitioning of different granularity local regions on distinct branches. Tu et al. [49] proposed a Multi-Granularity Mutual Learning Network (MMNet), where “multi-granularity” refers to adjusting the number of stripes in different branches to control the granularity of the learned features. Jiang et al. [50] proposed a dual-branch network for cross-modal person ReID, where one branch uses a butterfly-shaped attention module to learn fine-grained features, while the other branch utilizes a ResNet network to capture coarse-grained features. Notably, several works such as [6,51] used convolutional operations with different kernel sizes to produce diverse local clues for pedestrian images, which has been proven simple yet effective. In this paper, we follow this design and combine it with differentiated divisions of feature maps (pyramid structure) to explore the multi-granularity relationships between different parts and the entirety.

3. Proposed method

In this section, the process of the Multi-granularity Cross Transformer Network (MCTN) is presented. We first introduce the problem definition of person re-identification in Section 3.1. Then, the overall architecture of MCTN is described in Section 3.2. Next, the core components of MCTN will be discussed in detail, including the Multi-granularity Convolutional Layer (MCL) in Section 3.3, the Pyramidal Cross Transformer learning layer (PCT) in Section 3.4, and the Hierarchical Aggregation Strategy (HAS) in Section 3.5. Finally, we describe the loss function in Section 3.6.

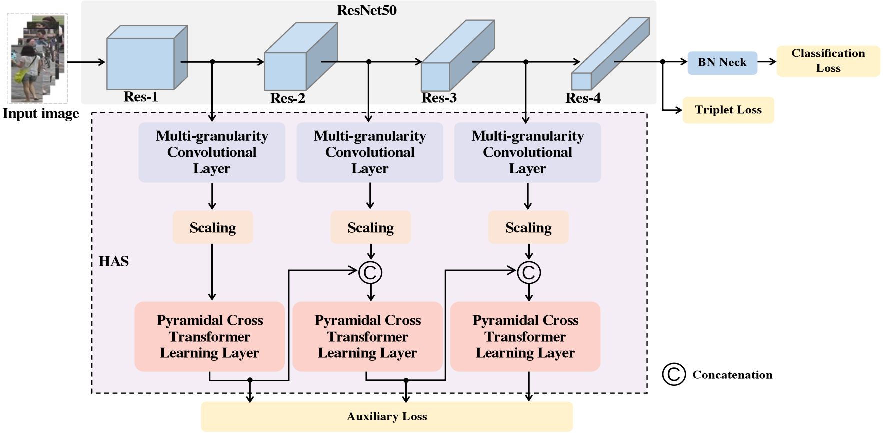

ResNet50
Input image
Res-1
Res-2
Res-3
Res-4
BN Neck
Classification Loss
Triplet Loss
Multi-granularity Convolutional Layer
Multi-granularity Convolutional Layer
Multi-granularity Convolutional Layer
Scaling
Scaling
Scaling
HAS
Pyramidal Cross Transformer Learning Layer
Pyramidal Cross Transformer Learning Layer
Pyramidal Cross Transformer Learning Layer
Ⓒ Concatenation
Auxiliary Loss

Fig. 2. The architecture of multi-granularity cross transformer network. Pedestrian images are first processed by ResNet50 to obtain global features. Then, the multi-granularity convolutional layer and pyramidal cross transformer learning layers are inserted in the first three stages of the backbone network. Hierarchical Aggregation Strategy (HAS) is employed to aggregate feature maps at different stages, mining more and richer local semantic information and improving the discrimination of the network. Finally, in the training phase, in addition to the loss of the backbone network, auxiliary losses are also added at each stage for supervised training. In the test phase, the features obtained by the backbone network and by HAS (the last stage) are combined as the final feature representation for test.

3.1. Problem definition

Given a training set $\mathcal{S}$ containing N pedestrian images, i.e., $\mathcal{S} = \{I_k, Y_k\}_{k=1}^{N}$, where I_k and Y_k are the kth pedestrian image and its corresponding identity. In the training phase, a network architecture $\mathcal{F}(\cdot)$ is designed to overcome the variations of pedestrians under different cameras to extract discriminative features, that is $f_k = \mathcal{F}(I_k)$, $f_k \in \mathbb{R}^{C\times H\times W}$, where C is channel number; H and W denote the height and width of the feature map, respectively. In the test phase, given a query image I_j, the trained network performs feature extraction for I_j and all pedestrian images in the gallery, obtaining discriminative feature maps f_j and $G = \{f_i\}_{i=1}^{M}$, where M is the number of pedestrian images in the gallery. The person re-identification results can be achieved by a similarity measurement between f_j and each element of G.

3.2. Overall architecture

The framework of MCTN is shown in Fig. 2. The process of the method mainly consists of five parts, including Backbone, Multi-granularity Convolutional Layer (MCL), Scaling, Pyramidal Cross Transformer Learning Layer (PCT), and Hierarchical Aggregation Strategy (HAS).

Backbone. Similar to previous studies, we utilize a ResNet50 model pre-trained on ImageNet as the standard backbone for extracting the global features of pedestrians. To bridge the domain gap between ImageNet and pedestrian image datasets, we also employ a ResNet50 model pre-trained on the LUPerson dataset, thereby enhancing the effectiveness of the extracted pedestrian features.

Multi-granularity Convolutional Layer. MCL is to simulate the process by which human vision recognizes things from different perspectives. More fine-grained features can be extracted through this layer to enhance the performance of the model.

Scaling. To reduce the parameters of the model and facilitate the integration of subsequent networks, Global Max Pooling (GMP) is applied after MCL. GMP can suppress the background information and extract significant features of pedestrians, making the salient features of pedestrians more compact.

Pyramidal Cross Transformer Learning Layer. The PCT is designed as a pyramid, guiding the network to discover salient features of pedestrians from a coarse-to-fine perspective. The attention features obtained with coarse-grained are further segmented for learning more fine-grained features. Thereby, complementary information of pedestrians with different granularities can be achieved.

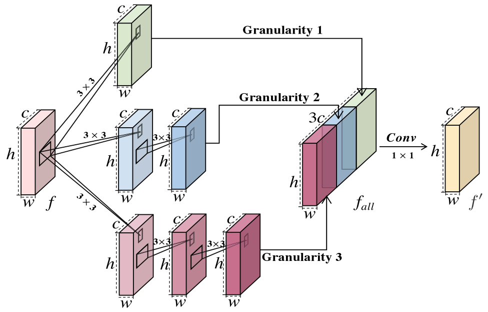


Fig. 3. The architecture of the multi-granularity convolutional layer. *Conv*: Convolutional operation; h, w, and c represent the height, width and channel number of the feature map, respectively. f, f' denote the input and output features, respectively. f_{all} indicates the features after concatenation.

Hierarchical Aggregation Strategy. HAS is used to obtain more comprehensive pedestrian features. Since the shallow layers contain more details, the features computed in the shallow layers can be aggregated into the deep layers, guiding them to pay more attention to fine-grained local features. Meanwhile, HAS is also helpful in mining the latent semantic information in the shallow layers by the interactions between the shallow layers and the deep ones.

3.3. Multi-granularity convolutional layer

As we know, not all discriminative features in pedestrian images can be obtained directly through the backbone network, even by continuously deepening the network. Intuitively, the human visual system usually observes things from different perspectives and granularities [28]. To this end, we introduce a multi-granularity convolutional layer and apply it to the first three stages of the backbone network to more comprehensively mine the semantic information contained in pedestrian images. The detailed structure of MCL is shown in Fig. 3.

Reviews. In Mudeep [19], the input data is further analyzed through three different receptive fields (a.k.a. granularity), i.e., 3 × 3, 5 × 5, and 7 × 7, respectively. The weights are not shared in the calculation process of feature representation at different granularities for obtaining richer features. To reduce the parameters while increasing the nonlinear transformation ability of the network, a 5 × 5 kernel is divided into two 3 × 3 cascades, and a 7 × 7 kernel is divided into three 3 × 3 cascades. The features extracted at different granularities

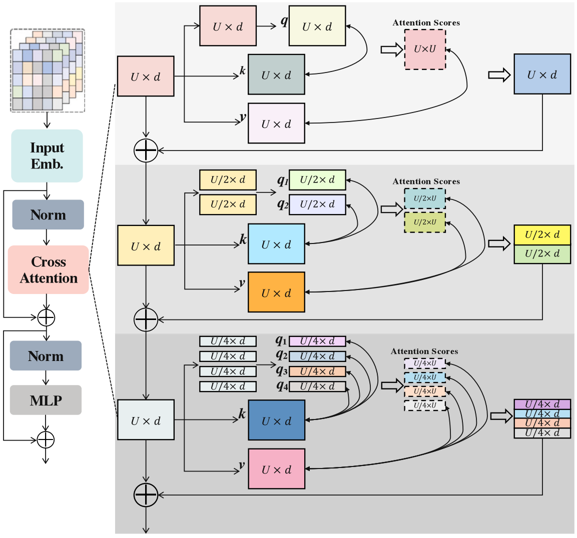


Fig. 4. The architecture of the pyramidal cross transformer learning layer. Blocks with different colors in the figure denote different linear transformations that are applied to the feature maps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

are fused as the final feature representation of this layer. Note that the convolution operations at different granularities are transformed into residual blocks. The input data goes through 1 × 1 convolution operations to compress the dimension of the feature, a 3 × 3 convolution calculations to extract the feature representation, and another a 1 × 1 convolution layer to restore the feature dimension to the original channel number. The aforementioned process is shown in Fig. 3. Finally, the currently obtained feature is connected with the output of the shortcut as the final feature representation. The above design has been proven to be effective.

Remarks. The MCL follows the above design since it has been proven to be effective. Although the structure of MCL is the same as Mudeep's multi-scale stream layer (MSL), very different usages and motivations are in this paper and [19]. Firstly, the features gained from MSL are used to guide the output features of each stream, and the features pass through an attention layer to filter out the redundant information, better distinguishing the pedestrian and the background. The MCL in this paper, however, is used to capture multi-scale local features for pedestrians, as the prerequisite of the subsequent pyramidal cross transformer learning layer to mine local information in the global structure from a coarse-to-fine perspective. Secondly, the MSL is only applied to the third layer of ResNet50, which acts as the fourth layer of the network structure. By contrast, we add MCL at the first three stages of ResNet50 to perform multi-granularity analysis on the input data to capture different feature representations. More comprehensive and discriminative features are acquired by fusing feature representations of different granularities.

3.4. Pyramidal cross transformer learning layer

Indeed, the performance of the Re-ID task largely depends on the semantic information of pedestrian images extracted by the neural network [15]. By deepening the network, the deep layers can, to some extent, learn the semantic information of pedestrian images. Nevertheless, the shallow layers contain only details and a small quantity of semantic information, and a lot of semantic information has not been mined yet. To this end, Transformer architecture based PCT is proposed to enrich the diversity of features. The detailed PCT structure is shown in Fig. 4. PCT mainly contains three components, including input embedding, cross attention, and channel MLP.

Input Embedding. Given a feature map X, it is first processed by input embedding, which is similar to patch embedding for ViTs [37]. The formulation can be expressed as:

$$X^{emb} = Norm(InputEmb(X)), \tag{1}$$

where $Norm(\cdot)$ denotes layer normalization [52], $X^{emb} \in \mathbb{R}^{U\times d}$ denotes the embedding tokens with sequence length U and embedding dimension d.

Cross Attention. To capture discriminative local information in the global structure from a coarse-to-fine perspective, we exploit the pyramid structure to explore the local–global relationship.

(1) Pyramid. The pyramid contains different levels of attention calculation, and each level horizontally splits the feature map into 2^{i-1}, where 2 is the radix and i is the current level. It is worth noticing that the granularity of horizontal divisions varies among pyramid levels. As the pyramid level increases, so does the division. The features learned at the coarse-grained level are used to guide the network to mine finer-grained local information. Additionally, considering that too fine a division will compromise the integrity of local semantic information, we set the pyramid level as 3.

(2) Cross. The feature map X_i^{emb} is horizontally split into m local feature vector X_{ij}^{emb}, where i represents the level of pyramid, $j = \{1, 2, \dots, m\}$ denotes the index of the local feature. The undivided feature map X_i^{emb} is transformed into K_i, V_i by two different linear transformations. To get more representative and discriminative features, we apply different linear transformations to different local vectors X_{ij}^{emb} producing distinctive features Q_{ij}. The feature calculated by cross attention can be defined as:

$$Y_{ij} = \sigma(Q_{ij}{K_i}^T/\sqrt{d})V_i + X_{ij}^{emb}, \tag{2}$$

where $K_i, V_i \in \mathbb{R}^{U\times d}$, $Q_{ij} \in \mathbb{R}^{U/m\times d}$ and K_i^T denotes the transpose of K_i. We investigate the relationship between local and global features using $Q_{ij} \times {K_i}^T$. Indeed, $Q_{ij} \times {K_i}^T$ is similar to the cosine similarity, so it can be utilized to measure the relevance between tokens. The softmax activation function $\sigma(\cdot)$ is used to normalize the obtained cross attention weights, and the local salient features in the global context are extracted by multiplying with V_i. Eventually, the obtained features are concatenated with the original local feature vector to gain cross attention feature representation.

(3) Merge. The feature maps obtained by each local attention are merged into the overall feature map at the current level, which can be formulated as follows:

$$Y_i = [Y_{i1}, Y_{i2}, \dots, Y_{im}], \tag{3}$$

where $[\cdot]$ denotes concatenating by height dimension. The feature map Y_i is then transmitted to the subsequent level of the pyramid, guiding the network to mine more fine-grained local semantic information in the global context. With the pyramidal cross transformer, the network becomes capable of effectively identifying the correlation between local and global features, highlighting the discriminative features in the local region and suppressing the irrelevant features.

Channel MLP. Channels of the feature map contain rich details and semantic information about the same parts of the pedestrian. Similar to [14,42], we retain the channel multi-layer perceptron (MLP) in the traditional transformer to condense the features of those similar channels. The channel MLP consists of two layers of linear transformation, Gelu activation function $\zeta(\cdot)$. Assuming that the feature map output by cross attention learning is Y, the procedure can be expressed as:

$$Z = \zeta(Norm(Y)W_1) \times W_2 + Y, \tag{4}$$

where Z denotes the output features of the Channel MLP and $W_1 \in \mathbb{R}^{d\times\tau d}$ and $W_2 \in \mathbb{R}^{\tau d\times d}$ are the learnable parameters with expansion ratio τ.

Remarks. Despite the fact that the architecture of the proposed PCT is similar to [14], they are totally different in detail. The Transformer structure in [14] is the original and plain one. In contrast, the PCT is specifically designed for the person Re-ID task, which includes the coarse-to-fine division of pedestrian image feature maps, differentiated feature extraction of different parts of pedestrians, and cross attention to exploring the local–global relationship of the feature map. Therefore, the PCT can obtain richer pedestrian features and improve the performance of the model.

3.5. Hierarchical aggregation strategy

Given the potential for shallow layers to contain untapped semantic information compared to deep layers [14], we employ a hierarchical aggregation strategy to mitigate this issue. Specifically, we simultaneously insert MCL and PCT into the first three stages of the backbone network. Notably, the feature maps obtained at each stage are not directly transmitted back to the backbone network for training, nor are they directly spliced together as the final feature representations. Instead, these features are fused with the features extracted by the MCL in the next stage, and inputted into the PCT to guide the network in discovering diverse clues of the pedestrian images. The aforementioned design is based on the following two reasons. Firstly, the independent processing can effectively and fully utilize the advantages of both CNN and Transformer, forming diverse local and global pedestrian features. Secondly, transmitting the features outputted by PCT to the backbone requires increasing dimensionality to ensure feature dimension consistency, which consequently significantly increases additional computational costs. In short, through the hierarchical aggregation strategy, the content across different levels can be more effectively utilized, thereby enabling the extraction of latent semantic information present in the pedestrian images.

3.6. Loss function

In the training phase, the verification loss, classification loss, and auxiliary loss are used to supervise the training of the multi-granularity cross transformer network.

Verification loss. The loss is used to enhance intra-class compactness and inter-class dissimilarity, i.e., to make the distance between positive and negative samples smaller by a predefined margin ξ. In other words, after the training, the distance between the same pedestrians will be as small as possible, whereas the distance between different pedestrians will be as large as possible. Similar to [14], we use the hard triplet loss [53] as the verification loss, which is defined as:

$$\mathcal{L}_{tri}(a,p,n) = [\|f_a - f_p\| - \|f_a - f_n\| + \xi]_+, \tag{5}$$

where f_a represents the anchor samples, f_p denotes the positive samples with the same identity, and f_n is the negative samples with different identities. $\|\cdot\|$ denotes the L_2-Norm, $[\cdot]_+$ indicates the max function.

Classification loss. Since the person ReID task can be regarded as an image classification problem with each pedestrian identity as a class, the classification loss used in this paper is cross-entropy loss, which is defined as:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{M} y_i log(\hat{y}_i), \tag{6}$$

where M is the number of pedestrian identities, $\hat{y}_i$ and y_i are the predicted label and ground-truth label of the ith pedestrian, respectively.

Auxiliary loss. To facilitate learning powerful feature representations of the network, the auxiliary loss $\mathcal{L}_{aux}$ consisting of verification loss and classification loss is added to different stages of HAS, which is defined as follows:

$$\mathcal{L}_{aux} = \sum_{i=1}^{S}(\mathcal{L}^i_{tri} + \mathcal{L}^i_{cls}), \tag{7}$$

where S is the number of stages. The auxiliary loss applied to the low and high levels of the network can fully exploit multi-granularity supervisions and enhance the interactions between the initial feature extraction of pedestrian images and the subsequent semantic information learning.

Finally, the total loss is:

$$\mathcal{L}_{all} = \mathcal{L}_{tri} + \mathcal{L}_{cls} + \gamma\mathcal{L}_{aux}, \tag{8}$$

where γ is a hyper-parameter of the network used to trade off the backbone network loss and auxiliary loss. For clarity, we show the detailed procedure of the proposed MCTN in Algorithm 1.

Algorithm 1 Multi-granularity Cross Transformer Network

```
Input: Training images $I$, epoch $T$
Output: Model $\mathcal{F}$
while $t <= T$ do
    Input $I$ into Res-1 block of ResNet50 (get $f_1$)
    Input $f_1$ into MCL (get $f_1'$)
    Scale $f_1'$ and implement PCT according to Eqs. (1)–(4) (get $Z_1$)
    for i = 2, 3 do
        Input $f_{i-1}$ into Res-$i$ (get $f_i$)
        Input $f_i$ into MCL (get $f_i'$)
        Scale $f_i'$ and concatenate $Z_{i-1}$ along the channel dimension (get $Z_{i-1}'$)
        Implement PCT for $Z_{i-1}'$ according to Eqs. (1)–(4) (get $Z_i$)
    end for
    Input $f_3$ into Res-4 block of ResNet50 (get $f_4$)
    Calculate the verification loss for $f_4$ according to Eq. (5)
    Perform BN Neck on $f_4$ and calculate the classification loss according to Eq. (6)
    Calculate the auxiliary losses for $Z_1$, $Z_2$, $Z_3$ according to Eq. (7)
    Calculate the total training loss to supervise the training of the model $\mathcal{F}$ according to Eq. (8)
end while
return the trained model $\mathcal{F}$
```

4. Experiments

In this section, we evaluate the performance of the proposed method on four large-scale person Re-ID benchmarks, including Market-1501 [54], DukeMTMC-reID [55], CUHK03 [56], and MSMT17 [57]. We first briefly introduce the information of these datasets and the commonly used evaluation metrics, then elaborate on the details of the implementation of our method. The effectiveness of each component of the method is demonstrated through extensive ablation studies. Qualitative visualizations of the attention map of our method compared with the baseline model HAT [14] are present. Quantitative comparisons with current state-of-the-art methods demonstrate the high performance and generality of the method. Finally, the robustness of the method is validated on the Occluded-DukeMTMC [58].

4.1. Datasets and settings

Market-1501 [54]. Market-1501 dataset was obtained in the campus market of Tsinghua University by 6 cameras, including 5 HD cameras and 1 SD camera. The dataset has a total of 32,668 images with 1501 identities, of which 12,936 images with 751 identities are used as the training set; 19,732 images with 750 identities are used as the test set. All images are produced by using the Deformable Part Model (DPM) [59] as the detector.

Table 1
The sizes of the feature maps for the input and output of each layer in our MCTN.

	Layer	Input size	Output size
stage1	Res-1	$B,\ 3,\ H,\ W$	$B,\ C,\ \frac{H}{4},\ \frac{W}{4}$
	MCL	$B,\ C,\ \frac{H}{4},\ \frac{W}{4}$	$B,\ C,\ \frac{H}{4},\ \frac{W}{4}$
	Scaling	$B,\ C,\ \frac{H}{4},\ \frac{W}{4}$	$B,\ C,\ \frac{H}{16},\ \frac{W}{16}$
	PCT	$B,\ C,\ \frac{H}{16},\ \frac{W}{16}$	$B,\ C,\ \frac{H}{16},\ \frac{W}{16}$
stage2	Res-2	$B,\ C,\ \frac{H}{4},\ \frac{W}{4}$	$B,\ 2C,\ \frac{H}{8},\ \frac{W}{8}$
	MCL	$B,\ 2C,\ \frac{H}{8},\ \frac{W}{8}$	$B,\ 2C,\ \frac{H}{8},\ \frac{W}{8}$
	Scaling	$B,\ 2C,\ \frac{H}{8},\ \frac{W}{8}$	$B,\ 2C,\ \frac{H}{16},\ \frac{W}{16}$
	PCT	$B,\ 3C,\ \frac{H}{16},\ \frac{W}{16}$	$B,\ 3C,\ \frac{H}{16},\ \frac{W}{16}$
stage3	Res-3	$B,\ 2C,\ \frac{H}{8},\ \frac{W}{8}$	$B,\ 4C,\ \frac{H}{16},\ \frac{W}{16}$
	MCL	$B,\ 4C,\ \frac{H}{16},\ \frac{W}{16}$	$B,\ 4C,\ \frac{H}{16},\ \frac{W}{16}$
	Scaling	$B,\ 4C,\ \frac{H}{16},\ \frac{W}{16}$	$B,\ 4C,\ \frac{H}{16},\ \frac{W}{16}$
	PCT	$B,\ 7C,\ \frac{H}{16},\ \frac{W}{16}$	$B,\ 7C,\ \frac{H}{16},\ \frac{W}{16}$
stage 4	Res-4	$B,\ 4C,\ \frac{H}{16},\ \frac{W}{16}$	$B,\ 8C,\ \frac{H}{16},\ \frac{W}{16}$

Table 2
Ablation analysis of components of MCTN on Market 1501 and DukeMTMC-reID. **Bold** indicates the best results.

Method	Market1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
-MCL -PCT -HAS -AL	84.7	93.5	76.4	86.8
-MCL	89.4	95.5	81.1	89.5
-PCT	87.6	94.6	78.4	87.8
-HAS	88.5	95.0	79.5	88.7
-AL	89.9	95.5	81.3	89.6
MCTN	**90.8**	**96.0**	**82.7**	**90.7**

Table 3
Ablation analysis of different number of PCT iterations on Market-1501 and DukeMTMC-reID. **Bold** indicates the best results.

Model	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
(1, 1, 1)	89.4	95.7	81.8	89.9
(2, 2, 2)	90.4	**96.0**	82.4	**90.9**
(3, 3, 3)	**90.8**	**96.0**	**82.7**	90.7
(4, 4, 4)	88.9	95.0	81.8	90.4
(1, 2, 3)	89.9	95.5	81.9	89.8
(2, 3, 4)	90.3	95.7	82.2	90.4
(3, 2, 1)	90.5	95.5	82.2	90.8
(4, 3, 2)	90.7	**96.0**	81.7	90.4

DukeMTMC-reID [55]. The DukeMTMC-reID dataset was captured by 8 HD cameras at Duke University. The dataset has a total of 36,411 images with 1404 identities, of which 16,522 images with 702 identities are used as the training set; 17,661 gallery images and 2228 query images with 702 identities are used as the test set.

CUHK03 [56]. The CUHK03 dataset was obtained at the Chinese University of Hong Kong by two of ten cameras. The dataset includes 13,164 images with 1467 identities. The images are produced in two ways: manually labeling the bounding boxes and a Deformable Part Model detector to automatically detect the bounding boxes. We partition the dataset according to the new protocol [60], where the training set contains 767 identities and the test set contains 700 identities.

MSMT17 [57]. The MSMT17 dataset is currently the largest person Re-ID dataset, captured by 12 outdoor cameras and 3 indoor cameras. The dataset contains 126,441 pedestrian images with 4101 identities. The training set contains 32,621 images with 1041 identities, and the test set contains 93,820 images with 3060 identities.

4.2. Evaluation metrics

Similar to most person Re-ID methods, this paper adopts mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) as the evaluation metrics of the person Re-ID methods, where mAP reflects the comprehensive performance of the Re-ID algorithm and CMC shows Re-ID accuracy by counting query identities among the top-N results.

4.3. Implementation details

The implementation details of our method can be described from three aspects: backbone model, preprocessing, and training settings.

Backbone. ResNet50 [61] is adopted as the backbone architecture for extracting global features of pedestrian images. To obtain a larger feature map, the stride of the last layer of the network is changed from the original 2 to 1. Unless otherwise specified, the pre-trained model is based on ImageNet.

Preprocessing. The images of all datasets are resized to 384 × 192. Horizontal flipping, random cropping, or erasing are adopted as data augmentation for training.

Training settings. The batch size is set to 64. Each batch randomly selects 16 person identities, each with 4 images. The max epoch is set to 150 and the Adam optimizer is adopted, of which the first 10 epochs use a linear warm-up strategy that the learning rate grows from 4×10^{-6} to 4×10^{-4}, and then decays by 0.4 every 20 epochs. Additionally, the triplet loss margin ξ is set to 0.3, and the hyper-parameter γ is set to 0.5 to balance the losses. All experiments are performed using PyTorch1.7 with 2 RTX A6000. In addition, for clarity, we show the sizes of the feature maps for the input and output of each layer in our MCTN in Table 1.

4.4. Ablation studies

Analysis of different components. To demonstrate the effectiveness of each component of the MCTN on the network, ablation experiments are performed on the Market1501 and DukeMTMC-reID by removing different components of MCTN. As shown in Table 2, "-MCL -PCT -HAS -AL" indicates that MCL, PCT, HAS, and AL are all not used. "-MCL" indicates that the feature maps obtained from the first three stages of the backbone network are directly used as the input of the pyramidal cross transformer learning layer without passing through multi-granularity convolutional layers. "-PCT" represents that the network aggregates the features obtained from the multi-granularity convolutional layers instead of the features from the pyramidal cross transformer learning layer. "-HAS" denotes that MCTN does not interact features obtained from shallow and deep layers. "-AL" indicates that MCTN is trained without using the auxiliary loss. From Table 2, we can see that, the mAP of MCTN drops by about 1.5% on both Market1501 and DukeMTMC-reID without multi-granularity convolutional layers, indicating that the MCL can render the network to extract richer pedestrian features. Without PCT or HAS makes the performance of MCTN significantly degrade on both Market1501 and DukeMTMC-reID. This demonstrates that the PCT can mine local information in the global structures from a coarse-to-fine perspective and the interactions between shallow layers and deep layers is helpful to learn rich features. Furthermore, supervised training of the network using auxiliary loss can also improve the performance.

Analysis of different numbers of PCT iterations. Beyond the pyramid level, we also conduct experiments to analyze the impact of different numbers of PCT iterations in the first three stages of the backbone network on performance. The same number of iterations in the first three stages of the backbone network is first conducted. The first four rows of Table 3 show that as the number of iterations increases, the network's performance improves. Two or three iterations can lead to better performance. However, when the iterations reach 4, the performance starts to degrade. This may be due to excessive

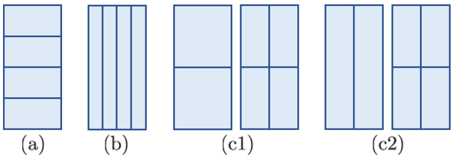


Fig. 5. Illustrations of different partitioning forms in PCT.

Table 4
Ablation analysis of pyramid levels on Market-1501 and DukeMTMC-reID. **Bold** indicates the best results.

Model	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
$\mathcal{P}_0$	87.6	94.6	78.4	87.8
$\mathcal{P}_1$	88.7	95.8	81.4	90.0
$\mathcal{P}_2$	89.7	95.9	81.8	90.2
$\mathcal{P}_3$	**90.8**	**96.0**	**82.7**	**90.7**
$\mathcal{P}_4$	89.3	95.7	81.8	90.6

Table 5
Ablation analysis of different partitioning sizes in PCT. **Bold** indicates the best results.

Method	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
1-2-4	**90.8**	**96.0**	**82.7**	**90.7**
1-3-5	89.0	95.3	81.5	90.3

Table 6
Ablation analysis of different partitioning forms in PCT. **Bold** indicates the best results.

Methods	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Ran-1
(a)	**90.8**	**96.0**	**82.7**	**90.7**
(b)	90.1	95.6	80.9	89.8
(c1)	89.9	95.5	80.7	89.7
(c2)	89.6	95.3	80.4	89.6

iterations, which lead to semantic information being lost in the deep layers and the network failing to obtain comprehensive features. Then, we discuss the effect of different iterations in shallow and deep layers on the performance. As shown in the last four rows of Table 3, for the performance on the Market-1501, (3, 2, 1) improves 0.6% mAP over (1, 2, 3). (4, 3, 2) improves 0.4% mAP and 0.3% Rank-1 over (2, 3, 4). On the DukeMTMC-reID, compared with (1, 2, 3), (3, 2, 1) increases 0.3% mAP and 1.0% Rank-1. However, (4, 3, 2) is lower than (2, 3, 4) by 0.5% mAP. In general, the performance of the network can be improved when there are more iterations in shallow layers than in deep layers, which also demonstrates that there is still some semantic information in shallow layers that needs to be mined deeply. Considering the generality of the network, all experiments are performed under the default setting (3, 3, 3).

Analysis of different pyramid levels. We report the performance of different pyramid levels in Table 4. $\mathcal{P}_0$ means that MCTN is trained without PCT. $\mathcal{P}_1$ represents that a PCT structure is only with one-level attention. $\mathcal{P}_2$, $\mathcal{P}_3$, $\mathcal{P}_4$ indicate that multiple levels of attention computation are utilized. Note that the granularity of horizontal divisions varies in different pyramid levels. As the number of pyramid levels increases, so does the quantity of local divisions. From Table 4, we can see that $\mathcal{P}_1$ outperforms $\mathcal{P}_0$ by 1.1%/1.2% and 3.0%/2.2% mAP/Rank-1 on Market-1501 and DukeMTMC-reID, respectively. This phenomenon shows that the pyramidal cross transformer learning layer enables the network to focus on diverse clues of pedestrians. The performance of the network gradually improves as the number of pyramid levels increases. The best performance of the model is achieved when the number of pyramid levels is 3 ($\mathcal{P}_3$), i.e., 90.8%/96.0% mAP/Rank-1 on Market-1501 and 82.7%/90.7% mAP/Rank-1 on DukeMTMC-reID. This validates that the features learned at the coarse-grained level can be used to guide the network in mining finer-grained semantic information. In addition, we observe that the performance of the network starts to degrade at a pyramid level of 4 ($\mathcal{P}_4$), which indicates that too fine a division of pedestrians may destroy the integrity of semantic information.

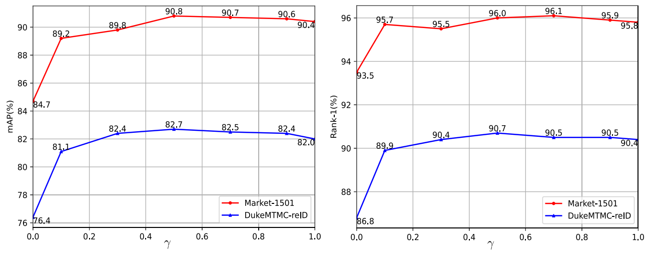
Analysis of different partitioning sizes in PCT. The partitioning sizes in PCT are essential to extract robust local features with semantics. We try two partitioning sizes, i.e., 1-2-4 and 1-3-5, and show the results in Table 5. It can be observed that using the partitioning scheme 1-2-4 achieves better performance compared to 1-3-5. This could be attributed to two factors. Firstly, the given pedestrian images are generally evenly distributed in terms of height dimension, so partitioning them into multiples of 2 can better preserve the semantic information of each local region. For example, pedestrian images can be divided into the upper body and lower body, and the upper body can further be divided into components such as the head and shoulders. Secondly, the 1-3-5 partitioning scheme results in smaller local regions, which may compromise the integrity of local semantic information. Therefore, we choose 1-2-4 as the default partitioning size.

Analysis of different partitioning forms in PCT. To validate the effectiveness of different partition methods in PCT, we try four ways, i.e., horizontal partitioning (a), vertical partitioning (b), and block-based partitioning (c1 and c2). Regarding block-based partitioning, there are two variations at the second level of the pyramid: one involves horizontal partitioning first (c1), and the other involves vertical partitioning first (c2). The illustrations of these four strategies are shown in Fig. 5.

From Table 6, it can be observed that vertical partitioning, compared to horizontal partitioning, resulted in a decrease in performance across all evaluation metrics on both datasets. This is mainly because vertical partitioning tends to destroy the integrity of local features in the human body while weakening their correlations. By contrast, horizontal partitioning can better preserve the semantic information of each local feature of pedestrians, such as the head, torso, and feet, thereby enhancing the performance of person re-identification. Regarding block-based partitioning, it achieves similar results to vertical partitioning and performs worse than horizontal partitioning in terms of mAP and Rank-1 metrics. This can be attributed to the fact that block-based partitioning also involves vertical partitioning, which results in the loss of semantic information. Therefore, we choose horizontal partitioning for preserving distinguishable pedestrian features.

Analysis of different pooling methods. To validate the effectiveness of Global Max Pooling (GMP) in the proposed method, we conduct the ablation study, i.e., replacing GMP with Global Average Pooling (GAP). From Table 7, it can be observed that on the Market-1501 and DukeMTMC-reID datasets, using GAP results in an average precision (mAP) drop of approximately 1.5% compared to GMP. This difference primarily stems from the fact that GAP fails to distinguish between pedestrians and the background. By averaging the information across the entire feature map, it may introduce background information unrelated to pedestrian identities, thus impacting the performance of person re-identification.

To further validate whether GAP/GMP will lead to the loss of fine-grained features, we conduct experiments by removing the GAP/GMP operation. Additionally, we adjust the stride size appropriately while fusing different granularity feature representations with multi-granularity convolution to ensure consistent feature map sizes across different stages. From Table 7, it can be observed that after removing the GMP operation, the model's performance on the Market-1501 and DukeMTMC-reID datasets decreased by 1.6% and 1.9% in mAP, respectively. It can be attributed to the presence of noise or redundant features in the original feature map, while the GMP operation directly extracts the most salient features, effectively capturing the local features of pedestrians.



(a)

(b)

Fig. 6. Performance statistics of the proposed MCTN with respect to different settings of hyper-parameter γ.

Table 7
Ablation analysis of different pooling methods. **Bold** indicates the best results.

Methods		Market-1501		DukeMTMC-reID	
GMP	GAP	mAP	Rank-1	mAP	Rank-1
✗	✗	89.2	95.7	80.8	89.7
✗	✔	89.5	95.7	81.0	89.8
✔	✗	**90.8**	**96.0**	**82.7**	**90.7**

Table 8
Ablation analysis of different granularities on Market-1501 and DukeMTMC-reID. **Bold** indicates the best results.

Model		Market-1501		DukeMTMC-reID	
		mAP	Rank-1	mAP	Rank-1
1	3 × 3	89.6	95.4	81.9	90.4
	5 × 5	90.4	95.4	82.1	90.6
	7 × 7	90.3	95.5	82.1	90.7
2	3 × 3, 5 × 5	90.4	95.6	82.4	90.4
	3 × 3, 7 × 7	90.6	95.7	82.4	**90.8**
	5 × 5, 7 × 7	90.5	95.8	82.2	90.3
3	3 × 3, 5 × 5, 7 × 7	**90.8**	**96.0**	**82.7**	90.7

Analysis of different granularities. We investigate the MCL further to determine how different granularities affect the performance of the network. The first three rows indicate that only one granularity is used for feature extraction. Three rows in the middle and one row at the bottom use two or three granularities, respectively. As shown in Table 8, the best performance can be achieved on Market-1501 and DukeMTMC-reID using three granularities. This suggests that simulating the human visual system to extract useful information from multiple perspectives can facilitate the network to extract rich and comprehensive pedestrian features.

Analysis of hyper-parameter γ. We discuss the impact of the hyper-parameter γ in the loss function on model performance. As shown in Fig. 6, we can see that the performance of the model on Market-1501 and DukeMTMC-reID increases first and then decreases with the increase of γ, and the optimal performance is achieved at $\gamma = 0.5$. $\gamma = 0$ means that MCTN is trained without using the auxiliary loss. This makes the training of feature maps in the middle of the network more difficult, leading to poor performance. $\gamma = 1$ indicates that the training of the MCTN depends equally on the backbone network and the HAS branch. The experimental results show that a good weighting of the backbone network loss and auxiliary loss will drive the model to obtain better performance. All experiments are performed under the default setting $\gamma = 0.5$.

4.5. Visualization

We show the attention maps of the baseline model HAT [14] and the proposed MCTN using smooth grad-cam++ [62]. It can be observed from Fig. 7(b) and (e), the heat of the attention maps of HAT is located in only a limited portion of the pedestrian image, which means that HAT only concentrates on the most salient features of pedestrian images. The heat of the attention maps of the proposed MCTN, on the other hand, is distributed across the entire pedestrian images as shown in Fig. 7(c) and (f). The phenomenon is mainly because the proposed MCL layer can produce diverse clues for pedestrian images and provide a good initialization for the subsequent PCT layer for further exploring the local–global relationships more sufficiently. Thus, the MCTN is capable of learning more comprehensive and discriminative pedestrian features, which is robust for many complex scenarios.

4.6. Comparisons with state-of-the-art methods

To validate the effectiveness of MCTN, we compare it with state-of-the-art methods on four widely used datasets, including Market-1501, DukeMTMC-reID, CUHK03, and MSMT17. We report mAP and Rank-1 metrics in Table 9. Due to the potential domain gap when applying the ResNet pre-trained on ImageNet to pedestrian re-identification tasks, which may not effectively reflect the model's performance, we also provide the results of our method using a ResNet50 pre-trained on the LUPerson dataset. LUPerson is an unlabeled dataset, consisting of 4M person images of over 200K identities. The images are captured in a wide range of environments, and thus have much better diversity. Subsequently, we will analyze and discuss the results for each of the four datasets respectively.

Market-1501. As shown in Table 9, without pretraining on the LUPerson dataset, our MCTN achieves 90.8% mAP and 96.0% Rank-1. The comparative methods fail to achieve a better balance between mAP and Rank-1 metrics. For example, NFormer achieves the best result in terms of mAP (91.1%), but relatively low Rank-1 (94.7%). Considering pretraining on the LUPerson dataset, MCTN* demonstrates the best performance, with 92.7% mAP and 96.9% Rank-1. This indicates that our model can better analyze input features with different granularities and capture more discriminative features through the novel multi-granularity learning layer and pyramidal cross transformer learning layer.

DukeMTMC-reID. Compared to the Market-1501 dataset, this dataset contains more occlusions and complex backgrounds, making the dataset more challenging. Table 9 shows that the proposed MCTN and APD [42] all achieve the best Rank-1 (90.7%), while as for mAP, MCTN outperforms APD by a substantial margin of 1.6%. Similar to the results on Market-1501, NFormer also achieves the best mAP (83.5%), but low Recall (89.4%). Our MCTN, pretraining on the LUPerson dataset, further improves the performance, obtaining 83.5% mAP and 91.6% Rank-1. This shows that our pyramidal cross transformer learning layer can flexibly extract discriminative local features in pedestrian images to deal with occlusions.

CUHK03. Fewer training samples on CUHK03 make it more difficult to develop a stable and effective model. As shown in Table 9, MCTN, without pretraining on LUPerson, attains the second-best results, achieving 81.5% mAP and 83.0% Rank-1 on Labeled and 76.9% mAP and 79.4% Rank-1 on Detected, respectively. PHA leads in performance, obtaining the highest mAP/Rank-1 (83.0%/84.5%), attributed to its effective strategy of enhancing high-frequency and reducing low-frequency ones. With pretraining on the LUPerson dataset, MCTN* outperforms others with 86.5% mAP and 87.8% Rank-1 on Labeled and 84.5% mAP and 86.4% Rank-1 on Detected, respectively. Compared with HAT [14], MCTN consistently improves both mAP and Rank-1 metrics, underlining the effectiveness of the novel components we introduced.

MSMT17. This dataset includes both indoor and outdoor images that were taken over a longer period of time and involved more illumination variations. It has a similar viewpoint to Market-1501, but much more complicated scenarios. As can be seen from Table 9, without considering pretraining on the LUPerson dataset, MCTN achieves the second-best performance with 66.4% mAP and 84.7% Rank-1. Fur-

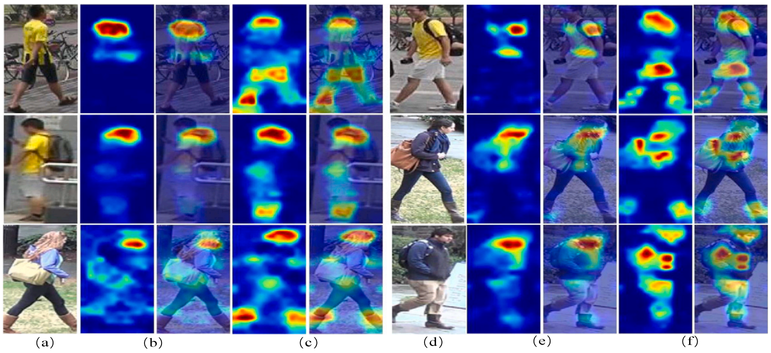


Fig. 7. Visualizations of 6 pedestrian images. (a) and (d) are input images. (b) and (e) are the visualizations of HAT [14]. (c) and (f) are the visualization of the proposed MCTN. The visualizations consist of an attention map and overlapped image.

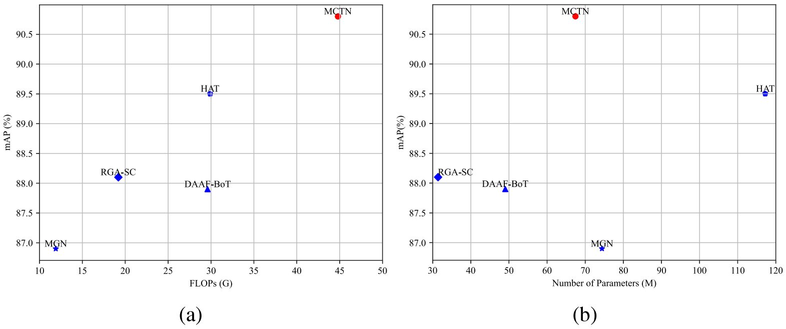


Fig. 8. Comparison of our method with other state-of-the-art methods in terms of FLOPs and parameters.

Table 9
Quantitative comparison with state-of-the-art methods on four Re-ID datasets, i.e., Market-1501 [54], DukeMTMC-reID [55], CUHK03 [56] and MSMT17 [57]. The comparative methods are cast into three categories, i.e, CNN-based methods, Transformer-based methods, and CNN+Transformer-based methods. **Bold** indicates the best results. "–" indicates the corresponding value is not provided in the original paper.

Methods	References	Market-1501		DukeMTMC-reID		CUHK03				MSMT17	
						Labeled		Detected			
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PCB [27]	ECCV18	81.6	93.8	69.2	83.3	–	–	57.5	63.7	40.4	68.2
MGN [28]	MM18	86.9	95.7	78.4	88.7	67.4	68.0	66.0	66.8	–	–
BOT [18]	TMM19	85.9	94.5	76.4	86.4	65.0	66.5	62.7	65.6	53.3	77.0
RGA-SC [7]	CVPR20	88.1	95.8	74.9	86.1	76.5	80.4	73.3	77.4	–	–
MEMF [20]	PR21	89.5	96.1	80.3	90.3	73.6	76.7	70.9	74.1	59.8	82.9
AGW [15]	TPAMI21	87.8	95.1	79.6	89.0	–	–	62.0	63.6	49.3	68.3
MMNet [49]	TITS22	88.7	95.3	80.3	89.7	–	–	–	–	–	–
DAAF-BoT [22]	PR22	87.9	95.1	77.9	87.9	67.6	69.0	63.1	64.9	–	–
LFS-ReID [63]	PR22	89.4	95.8	79.9	89.3	–	–	73.6	76.9	62.3	82.6
CGE-AGW [3]	PR23	90.1	95.6	–	–	78.1	79.8	75.0	77.4	64.5	83.9
CGE-AGW[a] [3]	PR23	92.2	96.3	–	–	86.4	87.4	84.2	85.9	65.9	85.1
ViT [37]	ICLR21	86.3	94.2	78.5	88.3	74.9	75.3	71.6	74.0	58.9	79.7
PAT [64]	CVPR21	88.0	95.4	78.2	88.8	–	–	–	–	–	–
TransReID [8]	ICCV21	88.2	95.0	80.6	89.6	–	–	–	–	64.9	83.3
DCAL [65]	CVPR22	87.5	94.7	80.1	89.0	–	–	–	–	64.0	83.1
PHA [10]	CVPR23	90.2	96.1	–	–	83.0	84.5	80.3	83.2	**68.9**	**86.1**
AAformer [66]	TNNLS23	88.0	95.4	80.9	90.1	79.0	80.3	77.2	78.1	65.6	84.4
HAT [14]	MM21	89.5	95.6	81.4	90.4	80.0	82.6	75.5	79.1	61.2	82.3
APD [42]	ICCVW21	89.1	95.8	81.1	90.7	77.2	79.9	75.3	78.1	61.2	82.4
FRT [12]	TIP22	88.1	95.5	81.7	90.5	–	–	–	–	–	–
NFormer [11]	CVPR22	91.1	94.7	**83.5**	89.4	78.0	77.2	74.7	77.3	59.8	77.3
MCTN (Ours)	–	90.8	96.0	82.7	90.7	81.5	83.0	76.9	79.4	66.4	84.7
MCTN* (Ours)	–	**92.7**	**96.9**	**83.5**	**91.6**	**86.5**	**87.8**	**84.5**	**86.4**	67.9	85.6

[a] Denotes using a ResNet-50 pre-trained on LUPerson [67].

thermore, although pretraining on the LUPerson dataset can improve performance, it also lags slightly behind PHA (mAP: 67.9% vs. 68.9%; Rank-1: 85.6% vs. 86.1%). The comparable performance of MCTN and MCTN* can also demonstrate the effectiveness of our method in handling complex conditions.

We also conduct a comparison of our method with other state-of-the-art methods in terms of FLOPs and parameters. From Fig. 8, it can be observed that our MCTN achieves the largest mAP and FLOPs among all competitors due to the introduction of a Multi-granularity Convolutional Layer (MCL) and a Pyramidal Cross Transformer learning layer

Table 10
Training and inference time of our MCTN.

Datasets	Total training time (h)	Inference time (ms)
Market-1501	4.25	2.03
DukeMTMC-reID	5.45	2.09

Table 11
Quantitative comparison with state-of-the-art methods on the Occluded-Duke and Occluded-ReID datasets. **Bold** indicates the best results.

Occluded-Duke			Occluded-ReID		
Methods	mAP	Rank-1	Methods	mAP	Rank-1
PCB [27]	33.7	42.6	PCB [27]	–	66.6
PVPM [68]	37.7	47.0	AFPB [69]	–	68.1
HOReID [70]	43.8	55.1	Teacher-S [71]	77.9	73.7
CBDB-Net [72]	38.9	50.9	REDA [73]	–	65.8
IGOAS [74]	49.4	**60.1**	IGOAS [74]	–	81.1
MCTN	**52.6**	59.1	MCTN	**80.1**	**87.4**

(PCT). Notably, our MCTN significantly improves performance while introducing certain parameters. Compared to the baseline HAT, we optimize the network, particularly with the introduction of PCT, which more effectively extracts salient features of pedestrians. This improvement allows for enhanced performance even with a reduced parameter count. Meanwhile, we show the total training time and per-image inference time of our MCTN on Market-1501 and DukeMTMC-reID in Table 10. It can be seen that our MCTN ensures reasonable training times while also delivering real-time processing capabilities, making it well-suited for applications that demand stringent immediacy.

4.7. Robustness analysis

To further verify the robustness of the MCTN, we conduct tests on the Occluded-Duke [58] and Occluded-ReID [69] datasets, which contain a significantly higher number of occluded images than the holistic pedestrian image datasets. As shown in Table 11, MCTN achieves the best mAP (52.6%) and the second-best Rank-1 (59.1%) on Occluded-Duke. Furthermore, it excels on the Occluded-ReID dataset, attaining top performance with 80.1% mAP and 87.4% Rank-1. Notably, IGOAS [74] is a method specifically designed for occlusion-based person Re-ID that uses a data augmentation method to produce occluded images for training, whereas the proposed MCTN is a general person Re-ID model. This demonstrates that our pyramidal cross transformer can eliminate the occlusion problem and mine local–global context from a coarse-to-fine perspective to enhance the robustness and generality of the model.

5. Conclusions

In this paper, we investigate how to design a dedicated Transformer for the person Re-ID task and combine its long-range modeling ability with CNN’s shift and scale invariance properties. To this end, we propose a network termed Multi-granularity Cross Transformer Network (MCTN). Specifically, the Multi-granularity Convolutional Layer (MCL) is used to produce diverse clues for person Re-ID at different granularities. We then devise a dedicated Transformer structure, called the Pyramidal Cross Transformer learning layer (PCT), to progressively learn local–global relationships from a coarse-to-fine perspective. To obtain more comprehensive and complementary pedestrian features, a Hierarchical Aggregation Strategy (HAS) is introduced to interact features learned by cross attention at different stages. The experimental results on four large-scale Re-ID datasets demonstrate the robustness and effectiveness of the proposed method.

Although the proposed method outperforms the state-of-the-art methods and obtains promising results on publicly available datasets, it still has something that can be improved in some aspects. (i) While the horizontal division of pedestrian images can generate local features for person re-identification, it lacks contextual understanding of the pedestrians. For instance, a component that encompasses complete semantics might be divided into different blocks. Therefore, human semantic parsing can be considered to replace the horizontal division, producing more discriminative local features. (ii) The MCTN is primarily designed to extract discriminative pedestrian features, yet it does not fully optimize the overall network structure. As a result, it may contain redundant information. Exploring network pruning methods as a means to streamline the network size and enhance inference speed presents a promising direction for future research.

CRediT authorship contribution statement

Yanping Li: Writing – review & editing, Writing – original draft. **Duoqian Miao:** Supervision. **Hongyun Zhang:** Validation. **Jie Zhou:** Visualization. **Cairong Zhao:** Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work is supported by the National Key Research and Development Program (Grant No. 2022YFB3104700), the National Natural Science Foundation of China (Grant Nos. 62376198, 62076182, 62163016, 62006172), the Jiangxi “Double Thousand Plan”, the Jiangxi Provincial Natural Science Fund (No. 20212ACB202001).

References

[1] Z. Li, D. Miao, Sequential end-to-end network for efficient person search, in: Proc. AAAI Conf. Artif. Intell., AAAI, Vol. 35, 2021, pp. 2011–2019.

[2] C. Zhao, Y. Tu, Z. Lai, F. Shen, H.T. Shen, D. Miao, Salience-guided iterative asymmetric mutual hashing for fast person re-identification, IEEE Trans. Image Process. 30 (2021) 7776–7789.

[3] J. Xi, J. Huang, S. Zheng, Q. Zhou, B. Schiele, X.-S. Hua, Q. Sun, Learning comprehensive global features in person re-identification: Ensuring discriminativeness of more local regions, Pattern Recognit. 134 (2023) 109068.

[4] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, D. Miao, Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification, IEEE Trans. Multimed. 22 (12) (2020) 3180–3195.

[5] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, ACM Trans. Multimed. Comput. Commun. Appl. 14 (1) (2017) 1–20.

[6] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, X. Xue, Multi-scale deep learning architectures for person re-identification, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, 2017, pp. 5399–5408.

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

[8] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2021, pp. 15013–15022.

[9] Z. Ye, C. Hong, Z. Zeng, W. Zhuang, Self-supervised person re-identification with channel-wise transformer, in: IEEE Int. Conf. Big Data, 2022, pp. 4210–4217.

[10] G. Zhang, Y. Zhang, T. Zhang, B. Li, S. Pu, PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2023, pp. 14133–14142.

[11] H. Wang, J. Shen, Y. Liu, Y. Gao, E. Gavves, Nformer: Robust person re-identification with neighbor transformer, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2022, pp. 7297–7307.

[12] B. Xu, L. He, J. Liang, Z. Sun, Learning feature recovery transformer for occluded person re-identification, IEEE Trans. Image Process. 31 (2022) 4651–4662.

[13] X. Liu, C. Yu, P. Zhang, H. Lu, Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–11.

[14] G. Zhang, P. Zhang, J. Qi, H. Lu, Hat: Hierarchical aggregation transformers for person re-identification, in: Proc. ACM Int. Conf. Multimedia, ACM MM, 2021, pp. 516–525.

[15] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2021) 2872–2893.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2016, pp. 2818–2826.

[17] X. Fan, W. Jiang, H. Luo, M. Fei, Spherereid: Deep hypersphere manifold embedding for person re-identification, J. Vis. Commun. Image Represent. 60 (2019) 51–58.

[18] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification, IEEE Trans. Multimed. 22 (10) (2019) 2597–2609.

[19] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, Leader-based multi-scale attention deep architecture for person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2019) 371–385.

[20] J. Sun, Y. Li, H. Chen, B. Zhang, J. Zhu, MEMF: Multi-level-attention embedding and multi-layer-feature fusion model for person re-identification, Pattern Recognit. 116 (2021) 107937.

[21] T. Si, F. He, H. Wu, Y. Duan, Spatial-driven features based on image dependencies for person re-identification, Pattern Recognit. 124 (2022) 108462.

[22] Y. Chen, H. Wang, X. Sun, B. Fan, C. Tang, H. Zeng, Deep attention aware feature learning for person re-Identification, Pattern Recognit. 126 (2022) 108567.

[23] Z. Zhang, Y. Wang, S. Liu, B. Xiao, T.S. Durrani, Cross-domain person re-identification using heterogeneous convolutional network, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1160–1171.

[24] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, 2017, pp. 1077–1085.

[25] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2018, pp. 1062–1071.

[26] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 7291–7299.

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

[28] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proc. ACM Int. Conf. Multimedia, ACM MM, 2018, pp. 274–282.

[29] B.N. Xia, Y. Gong, Y. Zhang, C. Poellabauer, Second-order non-local attention networks for person re-identification, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2019, pp. 3760–3769.

[30] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Deep attributes driven multi-camera person re-identification, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2016, pp. 475–491.

[31] X. Chang, T.M. Hospedales, T. Xiang, Multi-level factorisation net for person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2018, pp. 2109–2118.

[32] J. Lin, L. Ren, J. Lu, J. Feng, J. Zhou, Consistent-aware deep learning for person re-identification in a camera network, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 5771–5780.

[33] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proc. AAAI Conf. Artif. Intell., AAAI, Vol. 34, 2020, pp. 13001–13008.

[34] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, 2017, pp. 3754–3762.

[35] Z. Zhao, R. Song, Q. Zhang, P. Duan, Y. Zhang, JoT-GAN: A framework for jointly training GAN and person re-identification model, ACM Trans. Multimed. Comput. Commun. Appl. 18 (1s) (2022) 1–18.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. Adv. Neural Inf. Process. Syst., NIPS, Vol. 30, 2017.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: Proc. Int. Conf. Learn. Represent., ICLR, 2021, pp. 1–11.

[38] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Proc. Int. Conf. Mach. Learn., ICML, 2021, pp. 10347–10357.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2021, pp. 10012–10022.

[40] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: A survey, ACM Comput. Surv. 54 (10s) (2022) 1–41.

[41] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, C.-W. Lin, Structure-aware positional transformer for visible-infrared person re-identification, IEEE Trans. Image Process. 31 (2022) 2352–2364.

[42] S. Lai, Z. Chai, X. Wei, Transformer meets part model: Adaptive part division for person re-identification, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, ICCVW, 2021, pp. 4150–4157.

[43] C. Han, B. Jiang, J. Tang, Multi-granularity cross attention network for person re-identification, Multimedia Tools Appl. (2022) 1–19.

[44] C. Chen, M. Ye, M. Qi, J. Wu, Y. Liu, J. Jiang, Saliency and Granularity: Discovering temporal coherence for video-based person re-identification, IEEE Trans. Circuits Syst. Video Technol. 32 (9) (2022) 6100–6112.

[45] G. Zhang, J. Yang, Y. Zheng, Y. Wang, Y. Wu, S. Chen, Hybrid-attention guided network with multiple resolution features for person re-identification, Inform. Sci. 578 (2021) 525–538.

[46] X. Gong, Z. Yao, X. Li, Y. Fan, B. Luo, J. Fan, B. Lao, LAG-Net: Multi-granularity network for person re-identification via local attention system, IEEE Trans. Multimed. 24 (2021) 217–229.

[47] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2020, pp. 10407–10416.

[48] G. Wang, Y. Yuan, J. Li, S. Ge, X. Zhou, Receptive multi-granularity representation for person re-identification, IEEE Trans. Image Process. 29 (2020) 6096–6109.

[49] M. Tu, K. Zhu, Q. Miao, C. Zhao, G. Zhu, H. Qiao, G. Huang, M. Tang, J. Wang, Multi-granularity mutual learning network for object re-identification, IEEE Trans. Intell. Transp. Syst. 23 (9) (2022) 15178–15189.

[50] J. Jiang, K. Jin, M. Qi, Q. Wang, J. Wu, C. Chen, A cross-modal multi-granularity attention network for RGB-IR person re-identification, Neurocomputing 406 (2020) 59–67.

[51] Y. Yang, L. Jin, Multi-granularity feature fusion for person re-identification, in: Proc. Int. Conf. New Mater. Mach. Veh. Eng., Vol. 22, 2022, p. 101.

[52] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, arXiv preprint arXiv:1607.06450.

[53] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, 2017, arXiv preprint arXiv:1703.07737.

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

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

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

[57] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2018, pp. 79–88.

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

[59] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[60] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 1318–1327.

[61] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2016, pp. 770–778.

[62] D. Omeiza, S. Speakman, C. Cintas, K. Weldermariam, Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models, 2019, arXiv preprint arXiv:1908.01224.

[63] H. Gu, J. Li, G. Fu, M. Yue, J. Zhu, Loss function search for person re-identification, Pattern Recognit. 124 (2022) 108432.

[64] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, F. Wu, Diverse part discovery: Occluded person re-identification with part-aware transformer, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2021, pp. 2898–2907.

[65] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, Y. Shan, Dual cross-attention learning for fine-grained visual categorization and object re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2022, pp. 4692–4702.

[66] K. Zhu, H. Guo, S. Zhang, Y. Wang, J. Liu, J. Wang, M. Tang, AAformer: Auto-aligned transformer for person re-Identification, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–11.

[67] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li, D. Chen, Unsupervised pre-training for person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2021, pp. 14750–14759.

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

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

[70] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, J. Sun, High-order information matters: Learning relation and topology for occluded person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2020, pp. 6449–6458.

[71] J. Zhuo, J. Lai, P. Chen, A novel teacher-student learning framework for occluded person re-identification, 2019, arXiv preprint arXiv:1907.03253.

[72] H. Tan, X. Liu, Y. Bian, H. Wang, B. Yin, Incomplete descriptor mining with elastic loss for person re-identification, IEEE Trans. Circuits Syst. Video Technol. 32 (1) (2021) 160–171.

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

[74] C. Zhao, X. Lv, S. Dou, S. Zhang, J. Wu, L. Wang, Incremental generative occlusion adversarial suppression network for person ReID, IEEE Trans. Image Process. 30 (2021) 4212–4224.

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

Duoqian Miao is currently a Professor and a Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University. He serves as the President of the International Rough Set Society (IRSS), the Honorary Chair of the CAAI Granular Computing Knowledge Discovery Technical Committee, the Vice Director for MOE Key Lab of Embedded System & Service Computing, the Vice President of Shanghai Artificial Intelligence Society and Shanghai Computer Society. His interests include machine learning, data mining, big data analysis, granular computing, artificial intelligence, and text image processing. He has published more than 200 papers in IEEE Trans. Cybern., IEEE Trans. Inf. Forensic Secur., IEEE Trans. Knowl. Data Min., IEEE Trans. Fuzzy Syst., Pattern Recognit., Inf. Sci., Knowl-Based Syst. and so on. Representative awards include the Second Prize of Wuwenjun AI Science and Technology (2018), the First Prize of Natural Science of Chongqing (2010), the First Prize of Technical Invention of Shanghai (2009), the First Prize of Ministry of Education Science and Technology Progress Award (2007). He serves as an Associate Editor for Int. J. Approx. Reasoning, Inf. Sci. and Chinese Assoc. Artif. Int.

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

Jie Zhou received the Ph.D. degree from the Tongji University, in 2011. He is now an associate researcher in the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, China. From 2010 to 2011, he was a visiting scholar in the department of electrical and computer engineering at University of Alberta, Edmonton, Canada. From 2017 to 2018, he was an associate researcher in The Hong Kong Polytechnic University, Kowloon, Hong Kong. His current major research interests include uncertainty analysis, pattern recognition, data mining, and intelligent systems.

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