A Multi-Resolution Feature Fusion Method for Pedestrian Re-identification

Haitian Qin, Yang Xu, Xupeng Chen

Abstract—**Pedestrian re-identification technology enables accurate identification of individuals and is widely used in modern intelligent video surveillance systems to aid law enforcement, including criminal apprehension and locating missing persons. However, variations in lighting, background, resolution, and other imaging conditions captured by different cameras create significant challenges in pedestrian feature extraction, often leading to poor recognition accuracy. To overcome these challenges, this paper presents a Multi-resolution Feature Fusion (MRFF) method for pedestrian re-identification, based on the Pedestrian Re-identification Relational Network (RNFPR). This approach incorporates the Coordinate Attention (CA) module into the DenseNet161 network to enhance feature extraction capabilities. Improving the discriminative and recognition accuracy of features requires learning and fusing pedestrian features from multiple low-resolution images. This process enhances the expressive power of feature maps, ultimately improving pedestrian recognition performance. Additionally, this method introduces a multi-resolution feature fusion module that segments and integrates multi-resolution features from image data. This enables the model to effectively combine feature information from various resolution levels, resulting in a more comprehensive feature representation. Experimental results show that the MRFF method achieves a 1.3% increase in mean Average Precision (mAP) and a 1.0% improvement in Rank-1 accuracy on the Market1501 dataset. For the DukeMTMC-reID dataset, it provides a 0.2% increase in mAP and a 0.7% enhancement in Rank-1 accuracy. Consequently, the MRFF approach results in an overall mAP increase of 0.7% on the DukeMTMC-ReID dataset, significantly improving pedestrian gender re-identification accuracy.**

Index Terms—**Deep learning, Computer vision, Pedestrian re-identification, Multi-Dimensional Attention**

I. Introduction

Pedestrian re-identification is a computer vision task that leverages technology to track and recognize pedestrians, with the objective of achieving precise identification of individuals across diverse scenarios. This field is widely acknowledged as a critical subproblem in image retrieval [1-2].

However, pedestrian re-identification technology encounters various difficulties and challenges, including issues related to low image quality, occlusion, and the presence of similar identity features. Additionally, traditional methods are hampered by limitations such as inaccurate feature extraction and complexities in image matching, further exacerbating the challenges within this field. As a result, pedestrian re-identification has become an important and intricate topic of study within the field of computer vision [3-4].

Traditional pedestrian re-identification methods predominantly rely on a single feature extractor, which may not fully exploit the information within the image. In contrast, feature fusion methods can comprehensively describe the appearance and characteristics of pedestrians by integrating various levels and types of features, thereby enhancing recognition accuracy and robustness. Hyunjong Park et al. [5] introduced RNFPR (Pedestrian Re-identification Relationship Network), which integrates single block and other block features for enhanced performance. The authors introduced two modules, the one vs. - rest relationship module (ORM) and GCP. ORM utilizes the one-to-one correspondence between body parts, allowing each part-level feature to incorporate information about the respective part as well as other body parts. GCP extracts global feature maps from whole body parts by integrating GAP [6] and GMP methods.

This article presents an advanced Multi-Resolution Feature Fusion (MRFF) method built upon the Relational Network for Person Re-Identification (RNFPR). Distinct from the ResNet50 network utilized in reference [5], our approach employs DenseNet161 [7] for feature extraction. The DenseNet161 network is characterized by its dense connection architecture, which promotes effective feature fusion between layers, thereby enabling comprehensive information capture within the image. This architecture significantly enhances the feature expression capability and recognition accuracy. Additionally, DenseNet161 is designed with a reduced parameter count and optimized computational efficiency, which collectively contribute to its superior performance in pedestrian re-identification tasks.

Manuscript received March 29, 2024; revised July 25, 2024.

This work was supported by the National Natural Science Foundation of China (61775169), the Education Department of Liaoning Province (LJKZ0310), the Excellent Young Talents Program of Liaoning University of Science and Technology (2021YQ04).

H. T. Qin is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: 1824234130@qq.com).

Y. Xu is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (corresponding author, phone: 86-13889785726; e-mail: 705739580@qq.com).

X. P. Chen is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: 342014084@qq.com).

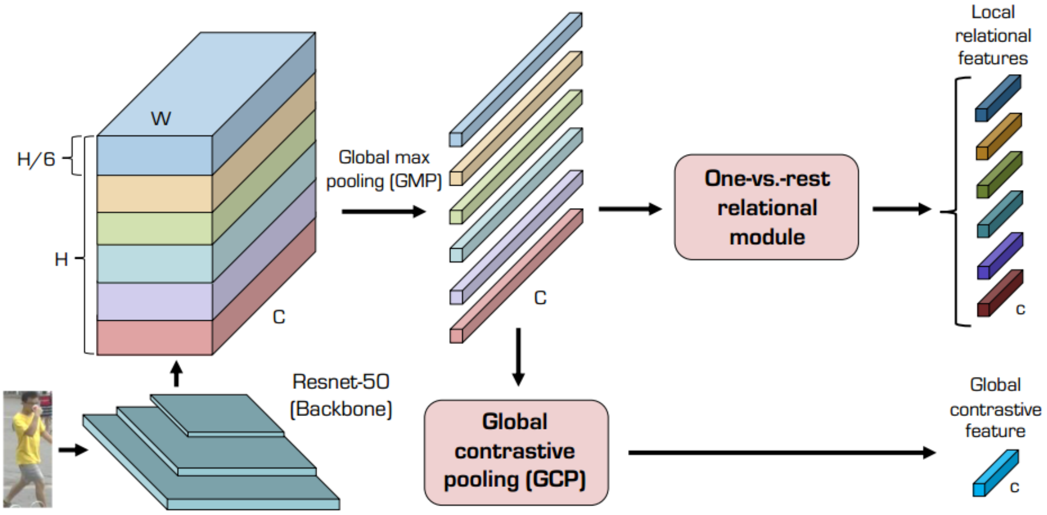
W
H/6
H
C
Global max pooling (GMP)
C
One-vs.-rest relational module
Local relational features
c
Resnet-50 (Backbone)
Global contrastive pooling (GCP)
Global contrastive feature
c

Fig. 1. Schematic Diagram of the RNFPR Algorithm

To further enhance the accuracy of pedestrian re-identification, a CA DenseNet161 model was constructed by integrating the attention module CoordinateAttention (CA) [8] into the first two modules of DenseNet161. This integration allows for improved focus on information from different regions in the image, thereby enhancing feature discrimination and robustness.

Moreover, this article presents a multi-resolution feature fusion module that decomposes the extracted features into multiple low-resolution features. This approach maximizes the utilization of effective feature information, mitigates interference from redundant data, and enhances the learning efficiency and performance of the model. By reducing reliance on large-scale features and enhancing the model's perception of small-scale features, overall accuracy is improved.

II. Principle of the RNFPR

RNFPR [5] introduced a pedestrian re-identification relationship network aimed at improving the accuracy and robustness of these tasks.

Pedestrian re-identification heavily depends on the precision and robustness of feature extraction and recognition, which are critical metrics in this domain. To address these challenges, RNFPR presents an innovative pedestrian re-identification relationship network. The algorithmic principle of RNFPR is depicted in Figure 1, which takes into account the relationships between individual body parts and other components. This approach facilitates the integration of partial information from various parts into a single partial level feature, thereby enhancing the model's discriminative power.

Within the RNFPR framework, the process begins with the use of the ResNet50 network to extract features from the input image. ResNet50 is a deep residual network engineered to improve classification accuracy by progressively increasing the network's depth. Subsequently, RNFPR segments the feature map into six parts along the H dimension and extracts part-level features using Global Max Pooling (GMP). Lastly, it computes local relationship features through Object Relation Modeling (ORM) and global contrast features via Global Context Pooling (GCP).

GCP is an advanced pooling method that harnesses contrast features to exploit variations in pooling results, thereby extracting complementary information and maximizing pooling features through residual learning. Additionally, it substantially improves the performance and accuracy of the overall model.

III. Improved Strategy

RNFPR utilizes ResNet50 as the feature extraction network, which incorporates residual connections. Notably, the input of each residual block is aggregated from the outputs of all preceding layers. However, this architectural design may lead to information loss and insensitivity to subtle features, thereby constraining its capability to extract complex features from pedestrian images. To address this challenge, the study utilizes DenseNet161 as a feature extraction network to extract features from images. This approach maximizes the utilization of features from previous layers, reinforcing feature transmission and reuse, consequently enhancing the network's capacity for feature expression and recognition accuracy. Additionally, this article incorporates the Channel Attention (CA) module into the DenseNet161 network, enabling the network to selectively focus on crucial features and thereby enhancing its feature extraction capabilities.

Therefore, the model proposed in this article demonstrates superior capability in extracting various features of different pedestrians within images, resulting in more accurate, discriminative, and robust feature representations. It successfully alleviates issues arising from cluttered backgrounds and low image quality, significantly boosting the accuracy of pedestrian re-identification. The MRFF approach detailed in this paper is comprised of three essential elements. Initially, the GridMask technique is applied for data augmentation to enhance the network's generalization performance. In the subsequent phase, the DenseNet161 network is utilized to replace the ResNet50 network. Leveraging dense connections, DenseNet enhances gradient backpropagation, rendering the network more trainable. Furthermore, its feature reuse and efficient calculations are achieved through short-circuit connections via feature connections, resulting in reduced parameters and improved computational efficiency.

Moreover, this article incorporates the Channel Attention (CA) module after the convolutional layers of the first two blocks of DenseNet161. This integration facilitates capturing inter-channel dependencies and modeling position information and long-range dependencies effectively, thereby significantly enhancing the model's feature extraction capability. In the third phase, this article introduces a multi-resolution feature fusion module that prioritizes local features. The extracted features are partitioned into multiple low-resolution features, with distinct local information emphasized through interval splitting. Ultimately, these features are fused to significantly enhance model accuracy while ensuring efficient learning and performance. By incorporating the CA DenseNet161 module along with the multi-resolution feature fusion module, the MRFF approach efficiently captures both local and global features, resulting in a notable enhancement in model accuracy.

A. *Optimization of Backbone Networks*

Neural networks play a critical role in machine learning, significantly impacting model performance. Therefore, the choice of neural networks directly affects the learning capability, generalization ability, and overall efficiency of models. In general, deeper networks tend to yield better performance. However, research has indicated that deeper networks can lead to decreased network performance, a phenomenon known as network degradation. This degradation is often attributed to the issue of gradient vanishing.

To address this issue, He et al. [9] proposed the ResNet architecture, which incorporates a robust mechanism that combines residual units to facilitate residual learning, effectively mitigating performance degradation in deeper networks [10]. DenseNet employs a dense connectivity strategy, where each layer's output is directly connected to the inputs of all subsequent layers. This strategy aims to optimize information flow between layers and enhance feature propagation. Specifically, in DenseNet, every layer is connected to all preceding layers via the channel dimension, thus providing input for the next layer.

Within the DenseNet framework, each layer is intricately linked to all previous layers through the channel dimension, and it serves as the input for the subsequent layer. This architectural design promotes extensive feature propagation throughout the entire network. In an L-layer network, DenseNet establishes a total of L（L+1）/2 connections. Additionally, DenseNet enhances the network by creating direct connections between feature maps of various layers, promoting effective feature reuse and significantly improving overall network efficiency.

Expressed in a formula, the output of the traditional network at layer l is formula 1 :

$$x_l = H_l(x_{l-1}) \tag{1}$$

In the ResNet architecture, the identity function derived from the preceding input is incorporated, as represented by formula 2:

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{2}$$

In DenseNet, all previous layers are connected as input, expressed as formula 3 :

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{3}$$

Within these operations, the aforementioned $H_l(\bullet)$ signifies nonlinear transformation, encompassing a combination of batch normalization (BN), ReLU, pooling, and Conv(3×3) operations. Additionally, several convolutional layers are situated between the l and $l-1$ layers. Analyzing the previously mentioned formulas reveals that the main difference between ResNet and DenseNet is in their connectivity strategies. ResNet employs skip connections to integrate residual blocks, whereas DenseNet creates dense connections by linking each layer's output to the inputs of all subsequent layers. The primary benefit of ResNets is their ability to facilitate direct gradient flow from deeper layers to shallower layers using the identity function. Nonetheless, the cumulative effect of combining identity mapping outputs with nonlinear transformations can, to some extent, impede the network's information flow.

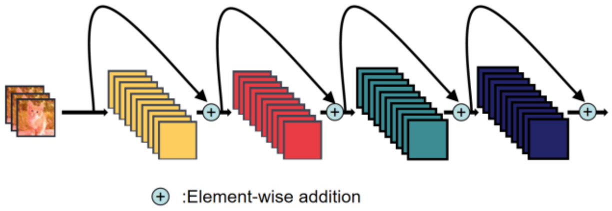

Fig. 2. Block structure of ResNet

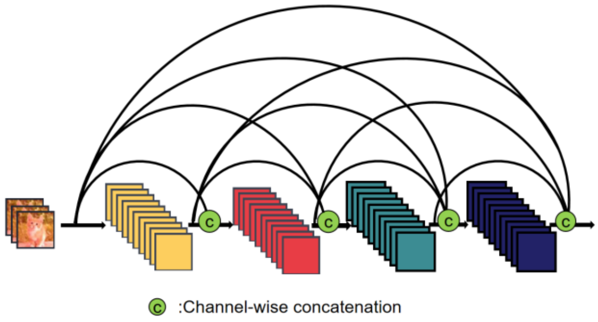

Fig. 3. Block structure of DenseNet

DenseNet establishes dense connections between the current layer and all subsequent layers, where $[x_0, x_1, ..., x_{l-1}]$ indicates the tensor concatenation of feature maps from layer 0 to layer $l-1$, effectively resulting in channel-wise superposition. This design facilitates the direct transmission of gradients through immediate connections to preceding layers, thus reducing the risk of gradient vanishing

In pedestrian re-identification, detailed image features are crucial due to the frequent similarity among dataset features. ResNet's architecture limits feature reuse as each residual block has restricted access to earlier layers' features. Conversely, DenseNet's dense connections enable more robust and comprehensive feature reuse.

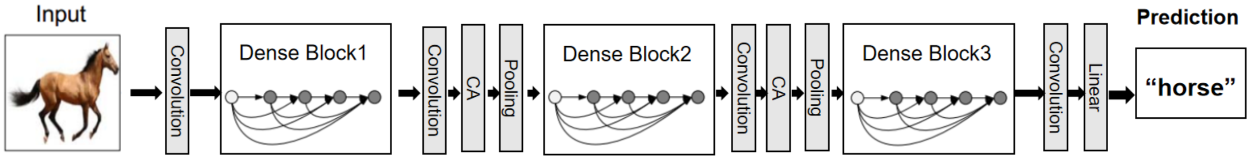

Fig. 4. Structure of CA-DenseNet161

It can effectively propagate and leverage fine-grained feature information within the image, thereby enhancing feature utilization efficiency and overall model performance. This feature is considered advantageous in pedestrian re-recognition models, such as RNFPR .

In the field of pedestrian re-identification, distinguishing various positions within an image is crucial for recognizing different pedestrian identities. To improve the accuracy of gender recognition among pedestrians, it is imperative to train a model that comprehends both global and local information in images. Consequently, this research incorporates the attention module [8] into the DenseNet feature extraction framework. The CA module acts as an attention mechanism, utilizing coordinate information to capture spatial relationships between different image positions effectively.

For pedestrian re-recognition tasks, this capability allows the model to focus on significant image regions, minimizing the impact of irrelevant details. In the DenseNet161 architecture, the module is incorporated after the first convolution layer of the initial two blocks. This addition enhances the extraction and use of spatial features, resulting in better training accuracy and overall model performance. Figure 4 depicts the structural diagram of the integrated feature extraction network CA-DenseNet161.

The Coordinate Attention (CA) module encodes both channel relationships and long-range dependencies by utilizing precise positional information. To enhance the attention module's ability to capture distant spatial interactions with detailed location information, global pooling is divided into two separate one-dimensional feature encoding processes, as demonstrated by formula 4:

$$k_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i, j) \tag{4}$$

Specifically, for a given input x, a pooling kernel with dimensions $(H,1)$ or $(1,W)$ is initially applied to encode each channel along the horizontal and vertical coordinates, respectively. Consequently, the output of channel c at height h is expressed by formula 5:

$$k_c^h(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h,i) \tag{5}$$

Similarly, the output of the c channel with width w is expressed by formula 6 :

$$k_c^w(w) = \frac{1}{H}\sum_{0\le j<H} x_c(j,w) \tag{6}$$

The two transformations consolidate features along two spatial axes, producing a set of direction-sensitive attention maps. These operations enable the attention module to identify targets with long-range dependencies in one spatial axis, thus improving the network's accuracy in target localization. To take advantage of the resulting representation, the two feature graphs generated above are concatenated and then transformed L_1 using a shared 1×1 convolution, expressed as formula 7 :

$$l = \delta(L_1([k^h, k^w])) \tag{7}$$

The tensor l is partitioned into two distinct tensors along the spatial dimension. The feature maps l^h and l^w are then adjusted to match the channel number of the input X through the application of two 1×1 convolutions, L_h and L_w , which are expressed as formulas 8 and 9 :

$$p^h = \sigma(L_h(l^h)) \tag{8}$$

$$p^w = \sigma(L_w(l^w)) \tag{9}$$

Extending p^h and p^w as attention weights, the final output of the CA module is expressed as formula 10 :

$$y_c(i, j) = x_c(i, j) \times p_c^h(i) \times p_c^w(j) \tag{10}$$

In the DenseNet161 network, the initial two blocks are designed to capture local features, while the CA module excels at linking local and global features. By incorporating the CA module with these initial blocks, the model's capacity to discern local features is notably enhanced. As the network progresses to subsequent layers, features tend to become more abstract and complex. Introducing CA modules at this stage may lead to an overly intricate attention mechanism, making it challenging to accurately capture the relationship between feature graphs. Consequently, this could potentially impact the model's performance. The integration of the CA module enhances feature propagation, improves the model's perceptual ability and understanding of image features, and aids in prioritizing important features while mitigating interference from insignificant information. Consequently, this contributes to enhancing the model's robustness and generalization capabilities.

B. Multi-resolution feature fusion module

In the realm of pedestrian recognition, advanced techniques such as SE attention [11] and CBAM [12] are highly effective in detecting essential image attributes. These attributes include key features like pedestrian body parts, the color and texture of clothing, background elements, and the spatial relationships among individuals. These features are essential for accurate pedestrian differentiation. Incorporating SE attention and CBAM into the recognition network improves the emphasis on these features, thereby enhancing both accuracy and robustness.

Nonetheless, these attention modules primarily augment feature map representational power through weight assignment. When the weights for certain features diminish toward zero, their impact on the final feature representation is significantly diminished or disregarded, which can lead to information loss and error propagation. As a result, these modules may fail to effectively capture the interrelationships among different features, limiting a comprehensive understanding of key object characteristics. Hence, feature fusion technology is crucial in computer vision [13], as it integrates channel, spatial, and target location features to preserve original data from various sources and avoid the loss of important information.

Following the extraction of initial feature maps from input images using the CA-DenseNet161 network, this study presents a novel multi-resolution feature fusion module, illustrated in Figure 6. After feature extraction, the resulting feature map is processed through the ShuffleAttention module before being fed into the multi-resolution feature fusion module. ShuffleAttention (SA) [14] utilizes the design principles of the SGE [15] attention mechanism, incorporates Channel Shuffle [16], and effectively merges spatial and channel attention mechanisms in parallel within the blocks.

The SA attention mechanism initially organizes the input feature maps into SA units, and subsequently divides each SA unit into two parts: one utilizing channel attention and the other employing spatial attention. This method efficiently combines both attention mechanisms within the SA units. Subsequently, the two parts of the SA are merged based on channel count to enable information fusion. Finally, all SA units are randomly integrated to produce the final output feature map. The integration of the SA attention mechanism enables the model to more effectively capture essential local information in pedestrian images, thereby enhancing recognition accuracy and robustness. Furthermore, by incorporating the SA attention mechanism, the model can prioritize important features during feature representation learning, reduce interference from redundant information, and ultimately improve overall learning efficiency and performance. The SA module structure is illustrated in Figure 5.

Next, the feature map output from the ShuffleAttention module is divided into multiple low-resolution segments, using 6 equal divisions as an example, and the input feature map is split at intervals based on (C,H,W). This strategy enhances the model's ability to capture multi-scale features in pedestrian images by leveraging low-resolution feature maps. It improves the model's processing speed and efficiency while retaining broader contextual information. Concurrently, this segmentation method serves to reduce the model's dependency on large-scale features, enhance its perception of small-scale features, and ultimately improve overall accuracy. The splitting diagram is illustrated in Figure 7, while the specific splitting formula is denoted by equations 11-16:

$$B_1 = (:,[1,4,7,10,...,H-2],[1,3,5,7,...,W-1]) \tag{11}$$

$$B_2 = (:,[1,4,7,10,...,H-2],[2,4,6,8,...,W]) \tag{12}$$

$$B_3 = (:,[2,5,8,11,...,H-1],[1,3,5,7,...,W-1]) \tag{13}$$

$$B_4 = (:,[2,5,8,11,...,H-1],[2,4,6,8,...,W]) \tag{14}$$

$$B_5 = (:,[3,6,9,12,...,H],[1,3,5,7,...,W-1]) \tag{15}$$

$$B_6 = (:,[3,6,9,12,...,H],[2,4,6,8,...,W]) \tag{16}$$

The low-resolution feature map B is divided according to (C,H,W), with features extracted from different parts. For optimal use, $H=3*n$, $W=2*m$, n and m should be natural numbers in the feature map dimensions. The fully connected layer mapping can also be applied to the dimensions where splitting occurs.

Upon the division into multiple low-resolution feature images, local features are obtained through the ORM module, while global features are extracted via the GCP module. The GCP module represents a hybrid method that integrates aspects of GAP[6] and GMP. Global Average Pooling (GAP) encompasses the entire body part of the image, but it is susceptible to interference from background clutter and occlusion. Global Max Pooling (GMP) partially addresses this issue by aggregating features from the most discriminative parts relevant to reID while disregarding background clutter. However, GMP does not encompass information from the entire body part. The GCP method facilitates the extraction of global feature mapping from the entire body, effectively mitigating interference from background and occlusion. As a result, this paper adopts the GCP method for extracting the global feature map.

Considering factors such as computational resources and practical requirements, alternative splitting strategies can be implemented. For example, the image can be partitioned into 4 or 2 segments by adjusting the sampling intervals accordingly. It is important to note that when the feature map is divided into fewer than 4 parts, the resolution of the subgraph after splitting will be relatively high. This may potentially compromise the efficacy of the multi-resolution feature fusion module. Therefore, it is recommended to strive for a higher number of subgraphs, ideally exceeding 4, in order to uphold the accuracy of the model.

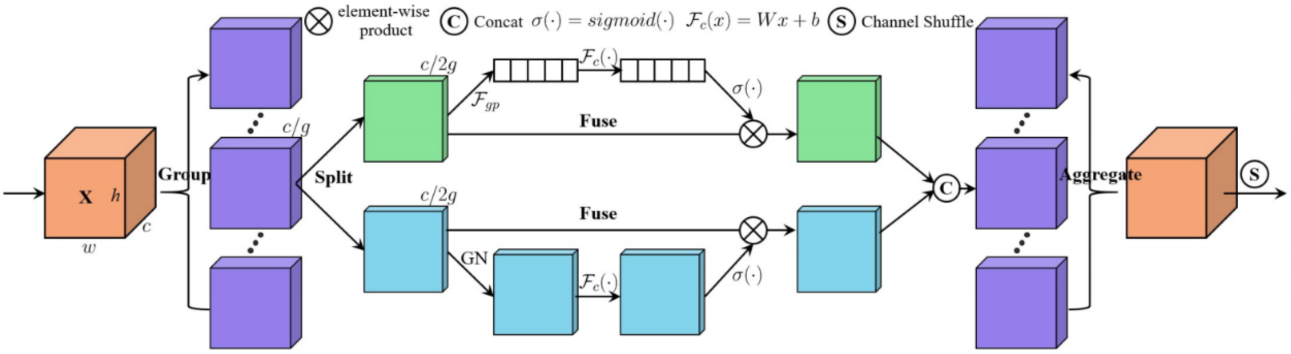

Fig. 5. SA module structure diagram

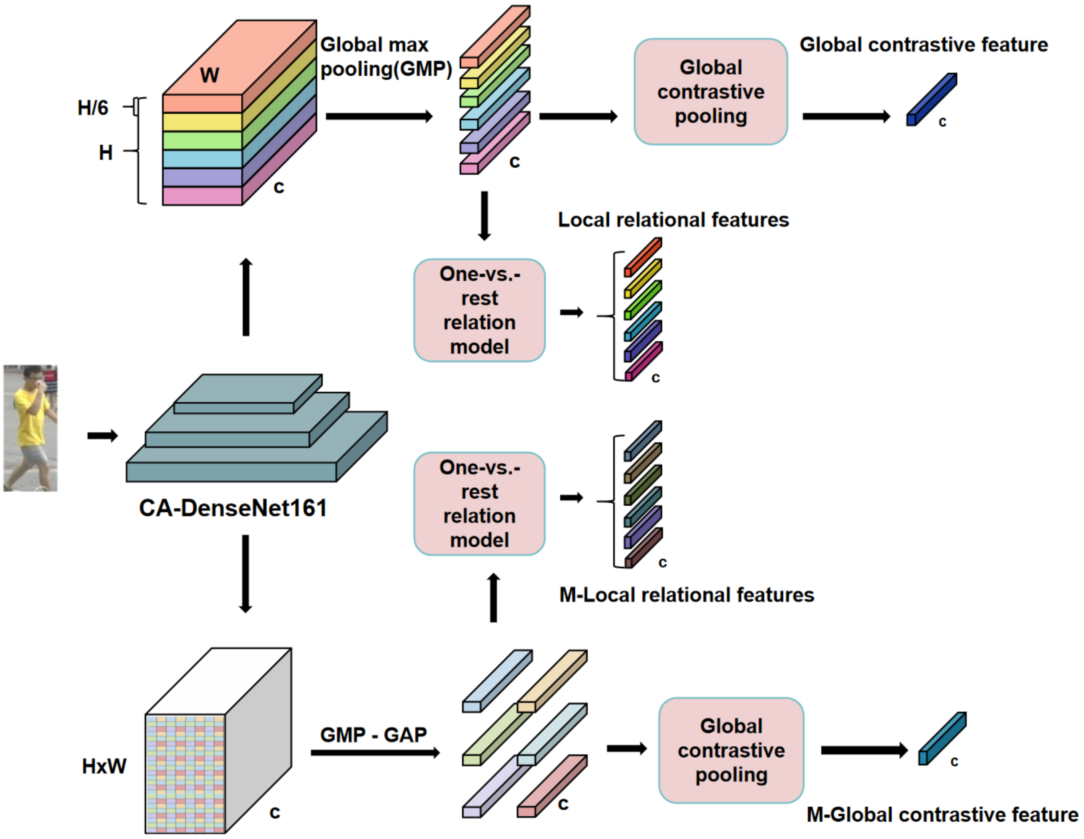
W
H/6
H
c
Global max pooling(GMP)
c
Global contrastive pooling
Global contrastive feature
c
Local relational features
One-vs.-rest relation model
c
CA-DenseNet161
One-vs.-rest relation model
c
M-Local relational features
HxW
c
GMP - GAP
c
Global contrastive pooling
c
M-Global contrastive feature

Fig. 6. Schematic diagram of MRFF method

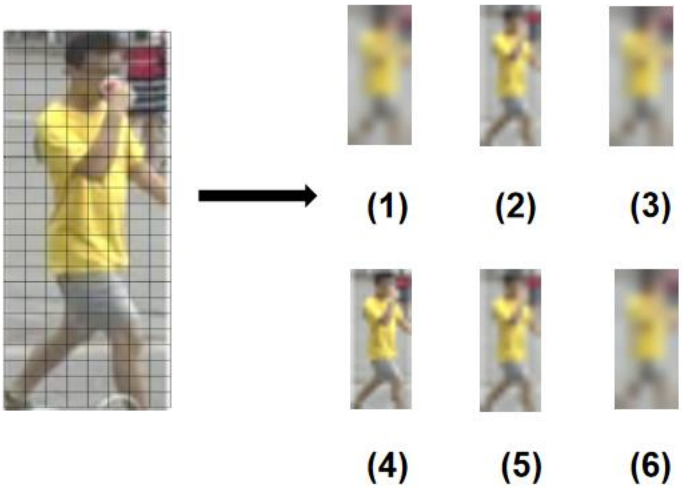

Fig. 7. Multi-resolution feature fusion module resolution diagram

IV. Analysis of Experiments And Results

The experimental platform is comprised of two key components: hardware and software. The hardware configuration features an Intel Core i7-11700 CPU paired with an NVIDIA GeForce GTX 3070 GPU. The software environment utilizes the PyTorch 1.8-GPU deep learning framework along with the PyCharm Community IDE for experiment development and training. A batch size of 64 was deliberately selected to maximize GPU efficiency, speed up training processes, and prevent problems related to memory overflow or suboptimal resource utilization. The feature extraction network's output channels are set to 2048 to ensure adequate information extraction and mitigate overfitting. Furthermore, configuring the number of subgraphs to 6 enhances the model's image comprehension capabilities and improves overall robustness. The number of hidden layer channels in the ORM module is established at 256 to ensure sufficient expressive capability, while also mitigating overfitting problems associated with excessive complexity.

A. *Training loss*

The training loss is a pivotal metric in deep learning, serving as an indicator of the variance between the model's predicted outcomes during training and the actual results. The primary objective of training loss is to minimize this disparity, thereby enabling the model to effectively acclimate to the training data and enhance its generalization capabilities.

Training loss is a crucial metric in deep learning, reflecting the difference between the model's predictions and actual outcomes during training. The main goal of training loss is to minimize this difference, thus allowing the model to better adapt to the training data and improve its generalization ability. In this study, the training loss L is composed of cross-entropy loss and triplet loss, balanced by the parameter λ as formula 17 :

$$L = L_{triplet} + \lambda L_{ce} \tag{17}$$

The triplet loss and cross entropy loss are expressed as $L_{triplet}$ and L_{ce} respectively, where the cross entropy loss is expressed as formula 18 :

$$L_{ce} = -\sum_{n=1}^{N}\sum_{i} y^n \log \hat{y}_i^n \tag{18}$$

Here, N denotes the number of images within the mini-batch, y^n represents the ground-truth identification labels, and $\hat{y}_i^n$ in indicates the predicted identification labels for each feature q_i . This is outlinedas formula 19 :

$$\hat{y}_i^n = \arg\max_{c \in K} \frac{\exp((w_i^c)^T q_i)}{\sum_{k=1}^{K} \exp((w_i^k)^T q_i)} . \tag{19}$$

K is the number of identification tags, w_i^k is the classifier of feature q_i and label k , using the fully connected layer as the classifier. In order to improve performance, batch-hard triplet loss[17] is used, which is expressed as formula 20 :

$$L_{triplet} = \sum_{k=1}^{N_K}\sum_{m=1}^{N_M}[\alpha + \max_{n=1\ldots M} \| q_{k,m}^A - q_{k,n}^P \|_2 - \min_{\substack{l=1\ldots K \\ n=1\ldots N \\ l \neq k}} \| q_{k,m}^A - q_{l.n}^N \|_2] \tag{20}$$

Here, N_K denotes the number of tags in the mini-batch, N_M represents the number of images per tag within the mini-batch, and $(N = N_K N_M) \cdot \alpha$ is a boundary parameter used to manage the separation of positive and negative pairs in the feature space. We employed $q_{i,j}^A , q_{i,j}^P , q_{i,j}^N$ to indicate anchor images, positive images, and negative images, respectively, where i and j refer to the identity index and image index respectively.

B. Dataset Selection

To address the diverse demands of pedestrian re-identification research, several relevant datasets are available. The experiments reported in this study utilized the Market1501 dataset [18], which comprises pedestrian images captured by six cameras installed across the Tsinghua University campus. This dataset includes annotated data for 1,501 individuals, encompassing a total of 32,668 images. The training set consists of 751 individuals and 12,936 images, while the test set includes 750 individuals and 19,732 images. This design maintains the independence of the data and the integrity of the evaluation process.

Due to its comprehensive annotations and complex scenes, this dataset has become a pivotal reference standard in pedestrian re-identification research. Representative examples from the Market1501 dataset are illustrated in Figure 8.

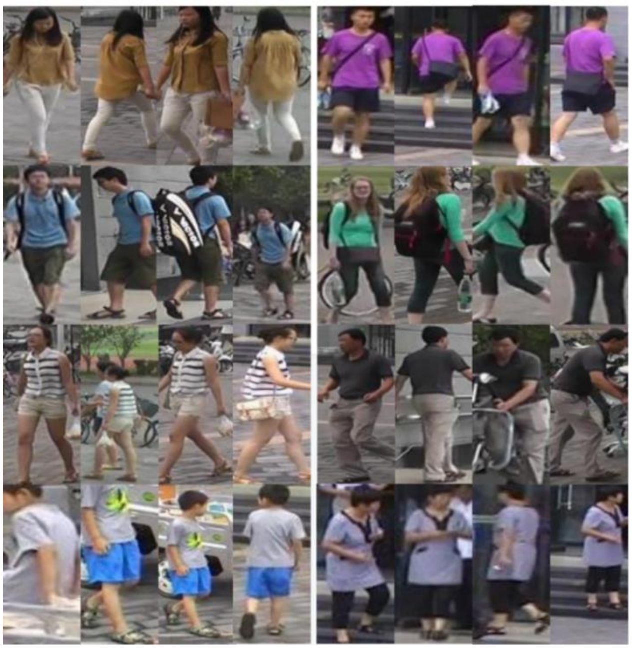

Fig. 8. Pedestrian pictures in the Market1501 dataset

C. Experimental Evaluation Criteria

This study employs Mean Average Precision (mAP) and Rank-n as evaluation metrics. mAP is a key metric for assessing a model's performance in retrieval tasks, combining the advantages of Precision and Recall. It reflects the average precision (AP) over multiple queries, thereby offering a comprehensive measure of retrieval accuracy. Conversely, Rank-n is crucial for evaluating retrieval systems, indicating the likelihood of finding the correct result within the top n search results. A higher Rank-n value reflects the model's improved ability to identify the correct result among the top n candidates. The mathematical formula for mAP is provided in Equation 21:

$$mAP = \frac{\sum_{i=1}^{m} AP_i}{m} \tag{21}$$

Where m represents the total number of categories, and AP_i denotes the average precision of the i -th category.

D. Experimental Analysis

Figures 9 and 10 illustrate the loss and accuracy curves for the ResNet50+RNFPR model, presenting performance metrics before and after implementing the proposed method. The analysis reveals that the enhanced method achieves a more rapid reduction in loss and a significant improvement in accuracy compared to the original approach. This outcome suggests that the proposed method substantially augments the model's ability to learn from crucial features, leading to faster convergence and a more stable and reliable training process.

By integrating the CA (Channel Attention) module into the DenseNet161 architecture, the model effectively captures inter-channel dependencies and models both spatial location information and long-range dependencies. This integration allows the model to discern and focus on the most relevant features, enhancing its ability to distinguish subtle differences between similar instances.

Following feature extraction, the multi-resolution features are segmented and fused within the newly implemented multi-resolution feature fusion (MRFF) module. This process facilitates the effective amalgamation of feature information across various resolution levels, yielding a more comprehensive and enriched feature representation. The MRFF module allows the model to utilize fine-grained details from high-resolution features and contextual information from low-resolution features, thereby improving its overall discriminative power.

Consequently, this approach significantly enhances the model's ability to focus on specific positional features crucial for accurate identification, thereby improving both accuracy and training stability. The refined feature integration supports the capture of critical details and fosters better generalization across different datasets. This enhanced capability is particularly beneficial in scenarios where precise identification is required, as it ensures that the model remains robust and reliable under varying conditions. The comprehensive feature representation achieved through this method provides a solid foundation for further advancements in model performance and application in real-world tasks.

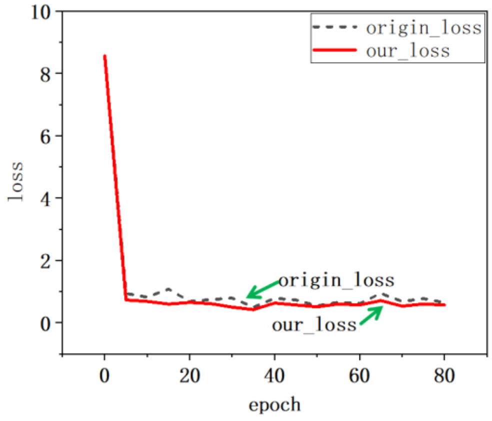

Fig. 9. Comparison curves of loss

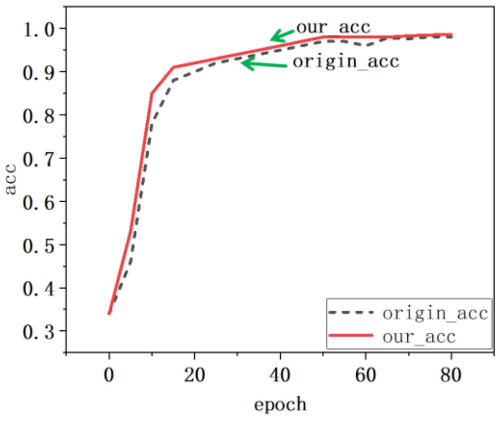

Fig. 10. Comparison curves of accuracy

1. Ablation experiment

To assess the effectiveness of the proposed enhanced algorithm, a series of ablation experiments were conducted, comparing various configurations: (1) the baseline ResNet50+RNFPR model, (2) utilizing CA-DenseNet161 for feature extraction, (3) incorporating a multi-resolution feature fusion module into the ResNet50+RNFPR model, and (4) implementing the MRFF method in the RNFPR model.

The experiments were carried out under controlled conditions using the Market1501 dataset, and the findings are summarized in Table I. The results show that replacing ResNet50 with CA-DenseNet161 for feature extraction increased mean Average Precision (mAP) by 0.8% and improved Rank-1 accuracy by 0.5%. Integrating the multi-resolution feature fusion module into the ResNet50+RNFPR model resulted in a 0.2% rise in mAP and a 0.3% gain in Rank-1 accuracy, with negligible effects on other performance metrics. In contrast, applying the MRFF method resulted in more significant improvements, with mAP increasing by 1.3% and Rank-1 accuracy by 1.0%. Moreover, this approach yielded notable gains in Rank-5 and Rank-10 metrics, further affirming the MRFF method's effectiveness.

2. Comparison experiment with mainstream algorithms

To evaluate the detection capabilities of the proposed algorithm, a comparative analysis was performed using established methods such as CAMA and BINet. This evaluation utilized the Market1501 and DukeMTMC-reID [19] datasets. Detailed results of this comparison are presented in Table II.

TABLE I
ABLATION EXPERIMENT

Model	mAP(%)	Rank-1(%)	Rank-5(%)	Rank-10(%)
ResNet50+RNFPR	88.9	95.2	98.2	98.9
CA-DenseNet161+RNFPR	89.7	95.7	98.4	99.1
ResNet50 + Multi resolution feature fusion module + RNFPR	89.1	95.5	98.3	99.0
MRFF+RNFPR	**90.2**	**96.2**	**98.5**	**99.1**

TABLE Ⅱ

COMPARED WITH ADVANCED ALGORITHMS

Model	Market1501		DukeMTMC-reID	
	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)
BDB [20]	94.3	84.5	86.9	72.3
UPR [21]	93.2	82.5	85.1	72.4
CBN [22]	94.3	83.6	84.8	70.1
CAMA [23]	95.2	84.7	86.1	73.2
BINet [24]	95.3	87.5	88.4	77.8
SORN [25]	95.2	84.7	87.1	74.2
LR-Net [26]	91.6	79.7	88.5	75.3
IGOAS [27]	93.6	84.5	87.2	75.4
HOReID [28]	93.9	84.5	86.9	75.6
OAMN [29]	93.5	80.1	86.2	72.6
BoT [30]	94.1	86.6	87.2	77.1
ResNet50+RNFPR [5]	95.2	88.9	89.7	78.6
MRFF(ours)	**96.2**	**90.2**	**89.9**	**79.3**

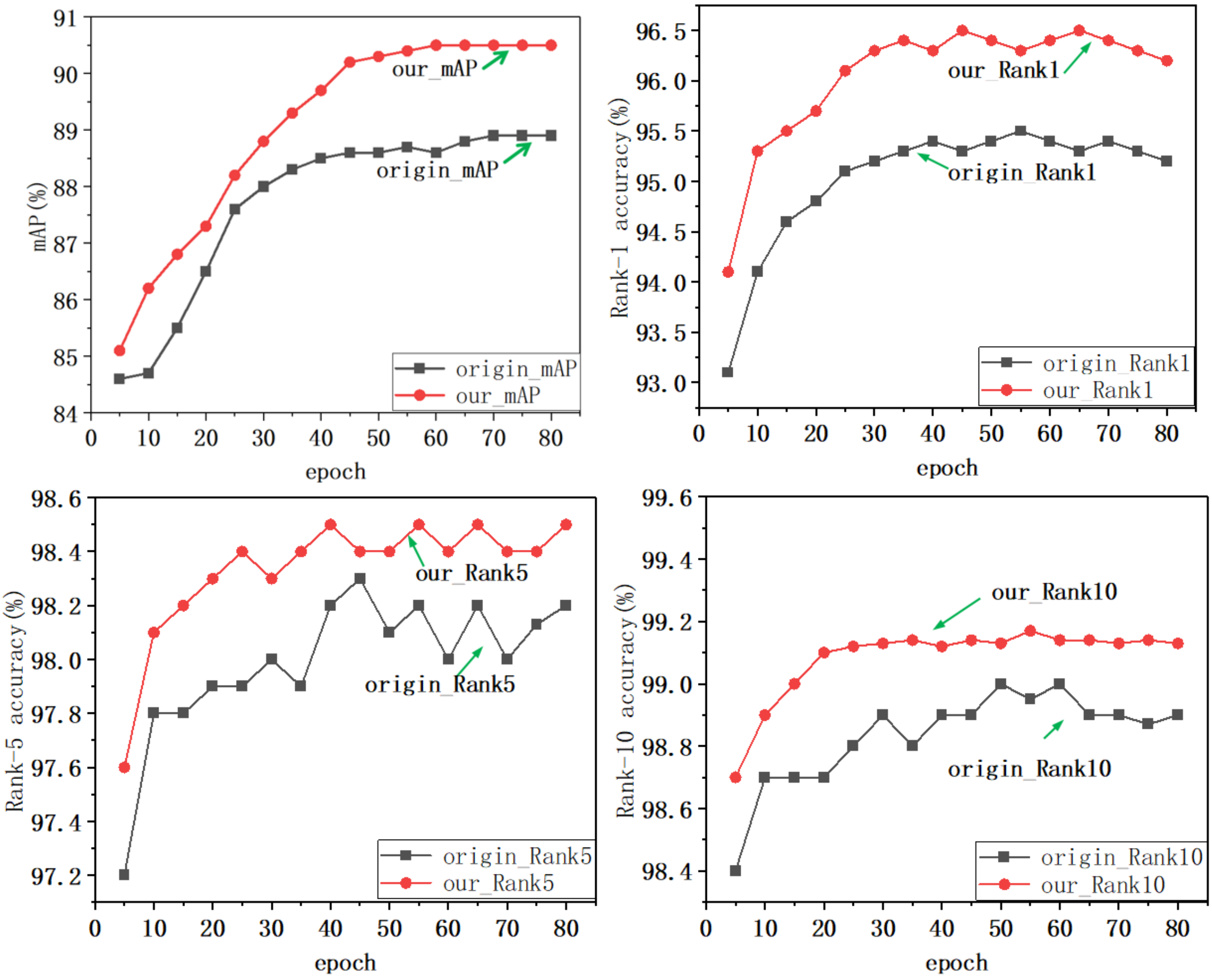

Fig. 11. Dot-line graph depicting changes in mAP and Rank-n metrics for the original and improved algorithms

The table data shows that our proposed method consistently outperforms other traditional algorithms. This suggests that the enhanced approach presented in this study surpasses existing methods, thereby validating the proposed algorithm's effectiveness. Furthermore, the DukeMTMC-reID dataset results show even better performance, highlighting the method's robustness and adaptability across various pedestrian re-identification datasets. Figure 10 visually depicts the trends in mean Average Precision (mAP) and Rank-n for both the ResNet50+RNFPR model and the proposed method, as shown in the line chart.

These results stem from 80 rounds of testing on the Market1501 dataset. It is important to note that the model's performance was significantly affected by the chosen parameter settings and hyperparameters during the experiments. Since extensive parameter optimization was not performed in this study, improvements in Rank-10 performance are somewhat limited. However, the enhanced algorithm demonstrates substantial gains in both improvement speed and stability compared to the original model. The RNFPR model utilizing this approach consistently achieved higher mAP and Rank-n scores than the baseline. The proposed method's learning process shows increased stability, resulting in more consistent and smoother performance. To further demonstrate this approach's effectiveness, pedestrian retrieval experiments were conducted using images from the Market1501 dataset.

V. CONCLUSION

This research proposes an RNFPR-based method for pedestrian re-identification, designed specifically to address issues such as high noise levels, poor image quality, and environmental factors that may affect model performance and reduce its sensitivity to fine details. The proposed methodology employs the CA-DenseNet161 network for feature extraction and integrates a tailored multi-resolution feature fusion module. This design enables the model to focus more effectively on pedestrian positional features through multi-resolution analysis, thus capturing fine-grained details with greater precision.

Consequently, the model's feature representation capability is significantly enhanced. Compared to the ResNet50+RNFPR model, the proposed method shows a 1.3% improvement in mean Average Precision (mAP) and a 1.0% increase in Rank-1 accuracy on the Market1501 dataset. Additionally, it demonstrates a 0.2% rise in mAP and a 0.7% enhancement in Rank-1 accuracy on the DukeMTMC-reID dataset. These findings suggest that the improved model offers superior accuracy and stability in pedestrian re-identification tasks. Future studies will aim to further refine this approach to address additional challenges in the pedestrian re-identification domain.

REFERENCES

[1] K. A. Shahrim, A. H. Abd Rahman, and S. Goudarzi, “Hazardous Human Activity Recognition in Hospital Environment Using Deep Learning,”*IAENG International Journal of Applied Mathematics*, vol.52, no.3, pp. 748-753, 2022.

[2] X. Zhang, M. Hou, X. Deng, and Z. Feng, “Multi-cascaded attention and overlapping part features network for person re-identification,”Signal, Image and Video Processing, vol.16, no.6, pp. 1525-1532, 2022.

[3] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, et al., “Deep learning-based methods for person re-identification: A comprehensive review,”Neurocomputing, vol.337, pp. 354-371, 2019.

[4] R. Zhu, Y. Xu, L. Wang, T. Sun, J. Yu, S. Ding, et al., “A Wide Range Multi-obstacle Detection Method Based on VIDAR and Active Binocular Vision,”IAENG International Journal of Applied Mathematics, vol.53, no.1, pp. 381-392, 2023.

[5] H. Park and B. Ham, "Relation Network for Person Re-Identification," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 11839-11847, 2020.

[6] M. Lin, Q. Chen, and S. Yan, "Network In Network," CoRR, abs/1312.4400, 2013.

[7] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269, 2017.

[8] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 13708-13717, 2021, doi: 10.1109/CVPR46437.2021.01350.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[10] J. Miao, S. Xu, B. Zou, and Y. Qiao, “ResNet based on feature-inspired gating strategy,”Multimedia Tools and Applications pp. 1-18, 2022.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 7132-7141, 2018, doi: 10.1109/CVPR.2018.00745.

[12] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional Block Attention Module," ArXiv, abs/1807.06521, 2018.

[13] M. Zhao, Q. Yue, D. Sun, and Y. Zhong, "Improved SwinTrack single target tracking algorithm based on spatio-temporal feature fusion," IET Image Process., vol. 17, pp. 2410-2421, 2023.

[14] Q.-L. Zhang and Y. Yang, "SA-Net: Shuffle Attention for Deep Convolutional Neural Networks," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2235-2239, 2021.

[15] X. Li, X. Hu, and J. Yang, "Spatial Group-wise Enhance: Improving Semantic Feature Learning in Convolutional Networks," ArXiv, abs/1905.09646, 2019.

[16] X. Wang, R. B. Girshick, A. K. Gupta, and K. He, "Non-local Neural Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794-7803, 2018.

[17] K. Zeng, M. Ning, Y. Wang and Y. Guo, "Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 13654-13662, doi: 10.1109/CVPR42600.2020.01367.

[18] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable Person Re-identification: A Benchmark," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1116-1124, doi: 10.1109/ICCV.2015.133.

[19] E. Ristani et al., "Performance Measures and a Data Set for Multi-target Multi-camera Tracking," ECCV Workshops, 2016.

[20] Dai Z, Chen M, Gu X, Zhu S, Tan P. Batch dropblock network for person re-identification and beyond[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 3691–3701.

[21] Liu T, Lin Y, Du B. Unsupervised person re-identification with stochastic training strategy[J]. IEEE Transactions on Image Processing, IEEE, 2022, 31: 4240–4250.

[22] Z. Zhuang et al., "Disassembling the Dataset: A Camera Alignment Mechanism for Multiple Tasks in Person Re-identification," ArXiv, abs/2001.08680, 2020.

[23] Shu X, Zhang L, Qi G-J, Liu W, Tang J. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2021, 44(6): 3300–3315.

[24] Chen X, Zheng X, Lu X. Bidirectional interaction network for person re-identification[J]. IEEE Transactions on Image Processing, IEEE, 2021, 30: 1935–1948.

[25] Zhang X, Yan Y, Xue J-H, Hua Y, Wang H. Semantic-aware occlusion-robust network for occluded person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, IEEE, 2020, 31(7): 2764–2778.

[26] Khan S U, Haq I U, Khan N, Muhammad K, Hijji M, Baik S W. Learning to rank: An intelligent system for person reidentification[J]. International Journal of Intelligent Systems, Wiley Online Library, 2022, 37(9): 5924–5948.

[27] Zhao C, Lv X, Dou S, Zhang S, Wu J, Wang L. Incremental generative occlusion adversarial suppression network for person reid[J]. IEEE

Transactions on Image Processing, IEEE, 2021, 30: 4212–4224.

[28] G. Wang et al., "High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6448-6457, 2020.

[29] P. Chen et al., "Occlude them all: Occlusion-aware attention network for occluded person re-ID," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11813-11822, 2021.

[30] H. Luo et al., "Bag of Tricks and a Strong Baseline for Deep Person Re-Identification," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1487-1495, 2019.