Person Re-Identification Algorithm Based on Improved ResNet

Wenrui Shen, Zhifeng Wang*

Abstract—**Person Re-Identification falls within the scope of computer vision, acting a technique to ascertain the presence of a specified pedestrian within a video or image library. The related research is of great significance in real-world environments such as criminal investigation and statistical analysis of commercial foot traffic and has received extensive attention from the academic community. However, traditional methods such as manual extraction cannot adapt to large-scale data volumes, and deep learning-based methods at this stage suffer from interference in complex environments such as similar costumes, perspective changes, and occlusion. Therefore, in this paper, we investigate the above problems. Firstly, we expand the dataset by introducing random erasure-based preprocessing of pedestrian images to enhancing the robustness and generalization capability of neural networks. Secondly, a composite attention mechanism is introduced after the network residual layer to enhance the spatial information capability and feature expression. Finally, the union loss composed of Circle Loss, Ternary Loss, and Cross Entropy Loss was chosen for network training in the loss optimization phase. Findings from the experiments reveal that the improved method proposed in this experiment achieves 96.0%Rank-1 and 88.3%mAP in Market1501, which reflects the validity of the approach proposed in this manuscript, and provides valuable reference suggestions for Person Re-Identification related research.**

Index Terms—**Person Re-Identification, Data Enhancement, Compound Attention Mechanism, Union loss function, ResNet.**

I. Introduction

As a pivotal component of modern smart video surveillance systems, Person Re-identification also called pedestrian re-identification, is a classical branch of the image retrieval problem [1]. Today it is widely used in criminal investigation, traffic analysis, and other fields. In addition, Person re-identification can also help mobile users achieve album clustering, assist retail or supermarket operators in obtaining effective customer trajectories, and explore commercial value. Person re-identification is a technique that determines whether a particular suspect pedestrian exists in an image or video against a predetermined pedestrian target, i.e., given a single monitored pedestrian image to retrieve a portrait of the pedestrian across multiple devices. This technique can make up for the visual limitations of the traditional fixed camera and can be merged with pedestrian detection and pedestrian tracking techniques, applied to video surveillance, intelligent security, and other areas [2], which has now become a trending issue in the field of computer vision. The Person Re-Identification task mainly consists of two key steps: feature extraction and similarity measure. Feature extraction is the first step after completing pedestrian detection, and its purpose is to cope with changes in pedestrians under different cameras, extracting representative features from input pedestrian images or video frames for subsequent comparison and recognition. Prior to 2014, Person Re-Identification relied heavily on manual feature extraction. Commonly used manual extraction methods are the histogram method and local feature-based methods, but the manual feature-based methods encounter performance bottlenecks when facing large-scale data. Originating from the continuous development of neural networks, methods based on deep learning are gradually proposed. Initially, scholars at home and abroad mainly used neural networks to learn the overall features of single-frame images [3], and this method can be categorized into representation learning and metric learning according to different types of loss, but the performance encountered a bottleneck. Then convolutional neural networks such as Inception, a module built from GoogLeNet [4], and the DenseNet family, which introduced dense connectivity, brought a series of breakthroughs. Traditionally the further the layers of a deep neural network, the richer the features obtained and the better the performance metrics obtained, but this is not the case. During the training process, conventional convolutional networks exhibit a degree of information loss and transfer depletion, along with the occurrence of the "gradient disappearance/explosion" phenomenon. As the network's depth escalates, accuracy saturates and then falls off sharply, a degradation usually caused by overfitting.

The ResNet network addresses this issue to some extent by concentrating solely on the disparities between the input and output. It achieves this by directly transmitting input information to output through the use of a bypass mechanism [5]. This approach ensures the preservation of information integrity and notably alleviates the problem of accuracy oversaturation that can arise as the network deepens. The ResNet50 network, which has been used in the realm of Person Re-Identification, and its most important feature is that it has 50 layers of convolutional neural networks, and advantage of depth of the network allows it to learn more intricate features, which enhances the precision while avoiding to cause gradient vanishing. In the ResNet network, if the inputs and outputs of a layer are equal, the layer is a constant mapping; if the inputs and outputs are different, the layer is a residual mapping. Residual learning is implemented through residual blocks, each containing two convolutional layers and a jump join so that the input is passed directly to

Manuscript received October 25, 2023; revised March 7, 2024. The research was backed by the National Natural Science Foundation of China (61575090, 61775169), the Natural Science Foundation of Liaoning Province (No.2019-ZD-0267) and the Liaoning Provincial Education Department (No.2020LNJC01).

Wenrui Shen is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail:1356452739@qq.com).

Zhifeng Wang is an associate Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author to provide phone: +086-150-4234-1839; e-mail:wangzhifeng_sia@126.com).

the output. However, the ResNet50 network also has some shortcomings. Firstly, it is prone to overfitting when there are not enough training samples. Secondly, the extracted pedestrian detail features are insufficient and the performance is not stable enough in the face of pedestrian pose changes and occlusions. Therefore, within this paper, we advocate an improved Person Re-Identification research method in light of ResNet50 network as a baseline model, introducing a random erasure preprocessing process, adding a composite attention mechanism, fusing the extracted pedestrian features with the image spatial pixel features, and finally optimizing the loss in the similarity metric stage, to propose a Person Re-identification research method based on the improved ResNet50, which get better performance in the ReID task.

II. Related work

Before deep learning gained prominence, research in Person Re-Identification relied heavily on traditional manual features such as color, texture, and physiological features of human body. Although these features are able to differentiate between different pedestrians in some cases, the feature extraction and labeling process involved is complex and often fails to capture robust pedestrian features. The initial study about Person Re-identification dates back to 1996 [6]. However, it was not until 2005 that the word "Person Re-Identification" appeared for the initial time in a paper in the realm of binocular vision tracking [7]. In following year, Person Re-id was first established as a separate branch of tasks within the field of computer vision, presented by Gheissari et al. at the CVPR conference [8]. Since then, related research has become a hot topic. The inaugural dataset, VIPeR, was publicly released by Gray et al. on the ECCV conference in 2008 [9]. Since then, the number of related papers and standard datasets began to increase rapidly. The inaugural symposium on Person Re-Identification took place on the ECCV international computer conference in 2012 [10].

The first professional book in this field was Person Re-identification, published in 2014 [11]. In the same year, with successful experience in the field of image classification, Li [12] et al. and Yi [13] et al. were the founder to integrate deep learning methods in Person Re-Identification. The exceptional performance has sparked widespread research interest, marking a new era in the advancement of Person Re-id through the application of deep learning methods. With the emergence and development of multiple neural networks, related research has gradually become deeper and broader, and the accuracy of their experiments on various datasets has been significantly improved. In 2012, AlexNet proposed by Hinton [14] and other scholars triumphed in the ImageNet Image Recognition Competition (ILSVRC), and in this regard, after the successful application of deep learning on image classification, it quickly spread to the field of Person Re-Identification [15]. When features are derived from the whole picture, they include not only human features but also background regions [16], but the background regions or occlusion may create misalignment problems. To address this challenge, literature [17] states that locating critical body parts and learning discriminative features can reduce interference and improve recognition accuracy. Shen et al. [18] proposed a novel compression-excitation network, SENet, which assigns weights to channels via feature compression and extraction. While obtaining better performance, current research on the attention mechanism continues to propose more complex modules. In order to reduce the model complexity, Wang et al. [19] suggested ECANet based on SENet, which significantly reduces the parameters while retaining performance. The SENet module determines the channel weights by means of two fully convolutional layers, whereas ECANet obtains the channel weights using a speedy one-dimensional convolution that is adaptively determined by a function related to channel. Chen et al. [20] integrated non-local attention and cyclic cross-attention into Person Re-Identification model. The non-local attention module boosts global feature extraction, while the cyclic cross-attention module simplifies complexity. Hou et al. [21] designed the attention mechanism by introducing a Bilateral Complementary Network (BiCent) formed by double branches. The initial branch uses the original image resolution. The subsequent subfield uses downsampled resolution to capture contextual details combined with a focus on the different physical characteristics of each identity. Ning et al. [22] proposed an attention mechanism for powerful salient features, which combines the acquisition of weakly salient features from images utilizing attention mechanism to achieve the extraction of global features, weakening some of the salient features in the process.

III. BUILD MODEL

A. Network Structure Design

The overall network structure design of Person Re-identification based on improved ResNet designed about this article is depicted in Figure 1. In the initial stage about feature extraction, the data undergoes augmentation through random erasure of the designated pedestrian present in the original surveillance video during model training. Subsequently, the backbone network of this paper was established by incorporating the composite attention mechanism module after each residual block in the following four phases of ResNet50, namely Conv2 x, Conv3 x, Conv4 x, and Conv5 x. The input picture passes through random erasure and enters Stage 0 to accept 64 convolutional kernels of size 7×7 convolutional operations before batch normalization and RELU activation function. Then, Channel Global Attention Mechanism RGA-C and the Spatial Global Attention Mechanism RGA-S are successively introduced in the residual layer after each stage of BottleNeck operation, as shown in the dotted box portion of Fig.1.

B. Random Erase Based Data Enhancement

The size of the data volume has a direct impact on the effectiveness of deep learning models. The phenomenon where a deep learning model excels on training data but falters on unseen test data, i.e., low training error and high test error, is commonly referred to as overfitting. Overfitting is typically due to insufficient amount of training data, resulting in a model that only remembers all the details in the training set, but is unable to generalize to new, unseen data. To avoid overfitting, it is crucial to ensure an ample size for the training sample. Nevertheless, in practice, various factors contribute to limited data collection. Therefore, there is a need to augment the number of training samples. Frequent

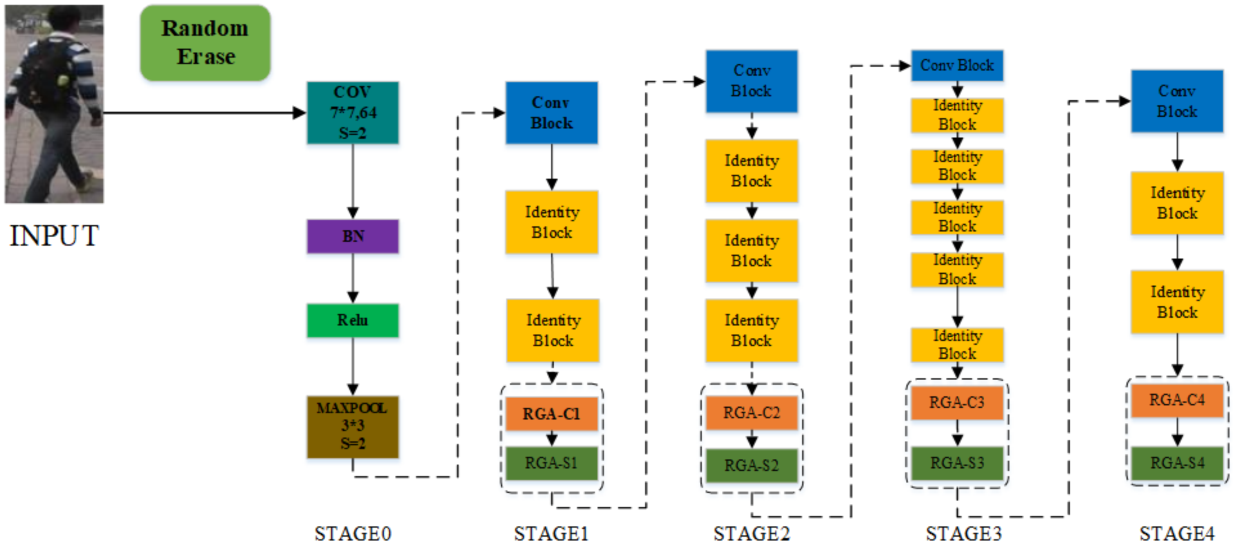

Fig. 1: Model Architectures

methods of data enhancement are random cropping and random flipping. As the network structure in deep learning continues to evolve, becoming increasingly intricate and profound, conventional data enhancement methods prove insufficient to meet the demands of Person Re-Identification application scenarios. Considering the diverse range of shooting backgrounds and angles, new challenges arise for the network. In this study, we employ random erasure generation to simulate occlusion phenomena in Person Re-Identification. This approach serves to broaden the dataset, enhancing the robustness and generalization capabilities of the deep learning network. The specific algorithm is as such:

Let W be the width of the picture and H be the height of the picture, then the overall image area S:

$$S = W \times H \tag{1}$$

Preset the maximum erase area length is A, the minimum erase area length is a, they satisfy the equation (2) :

$$0 < a < A \leq min(W, H) \tag{2}$$

1-5 selected regions are randomly generated in the image, and each randomly generated erased region satisfies the following conditions. Where (x_1,y_1,x_2,y_2) is each randomly selected region.

$$\begin{aligned} x_1 &= rand(0, W - A) \\ y_1 &= rand(0, H - A) \\ x_2 &= x_1 + rand(a, A) \\ y_2 &= y_1 + rand(a, A) \end{aligned} \tag{3}$$

Some selected rectangular erased areas have a pixel value of 255. The preprocessed image is shown in Figure 2.

Fig. 2: Data Enhancement Effects

C. Composite Attention mechanism

Attentional mechanisms derive from brain signal processing mechanisms specific to vision by humans. Humans typically scan the entire image broadly to identify the target area of interest. They then concentrate more attention on this specific area to gather detailed information about it. Factors such as background, occlusion, and significant changes in pedestrian pose interfere with the Person Re-Identification problem and can significantly affect the model results. For the model to pay better attention to the focus region in the image based on an in-depth study of the attention mechanism, a fusion attention mechanism is proposed to enhance the features extracted from a specific region by introducing a high weight to obtain more robust features, which in turn counteracts the adverse effects of interfering information.

Regarding this study, we design a Composite Attention Mechanism module that sequentially adds Channel Global Attention and Spatial Global Attention following residual layer of the backbone feature extraction network ResNet50. In the Channel Global Attention mechanism, the quantity of feature channels extracted from the pedestrian picture after processing through the baseline network is 1024. However, among these 1024 channels, numerous interfering channel features exist. Thus, it becomes imperative to filter these channel features and extract the key channels. Convolutional kernel in spatial global attention mechanism generates feature maps after continuously sliding over the pedestrian image. This process finely extracts both valid and invalid features, effectively filtering out the extraneous and useless features. Introduce a composite attention mechanism module after each Res layer in Stage 1 to Stage 4 in the baseline model ResNet50. The configuration of each composition attention mechanism module depicted in Figure 3, the feature map of the target pedestrian image to be queried is F, and the output is F_c,F_{cs} after passing through the channel global attention module and the spatial global attention module, respectively.

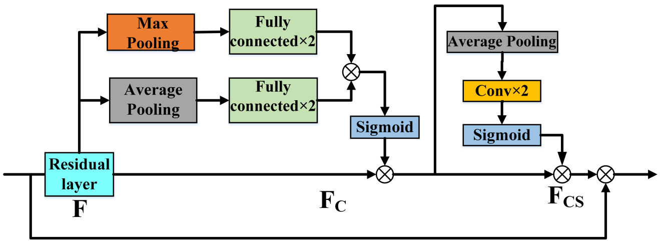

Fig. 3: Composite Attention Mechanism Structure

1) Channel Global Attention: Figure 4 illustrates the Channel Global Attention Mechanism, which operates by learning the weights of each channel to accentuate the key channels within them, ultimately yielding a one-dimensional channel attention weight, Mc. The structural framework details of the Channel's Global Attention mechanism are outlined below:

A feature map F with height H, width W, and number of channels C is simultaneously input to the Global Average Pooling (GAP) and Global Max Pooling (GMP) layers to compute the average and maximum values of all the channels to obtain two features of size 1×1×C.

After passing through two Fully Connected layers (FC), the weights for each channel are determined. To minimize parameter count and simplify the module, the hidden layer of the first Fully Connected layer contains C/r neurons, where r is the compression ratio. Meanwhile, the hidden layer of the second Fully Connected layer comprises C neurons, thereby reinstating the size to its initial specification. Following the channel attention link, the width (W) and height (H) of the feature map remain unchanged, preserving relevant information from different channels.

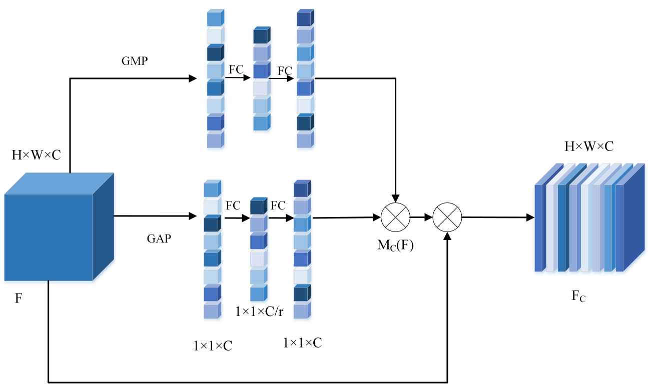

Fig. 4: Channel Attention Mechanism Module Framework

Applying the channel attention weights to the feature map outputs the channel attention feature map F_c with the following equation:

$$\begin{aligned}\mathrm{F_C} &= \mathrm{F} \times \mathrm{M_C(F)} \\ &= \mathrm{F} \times \sigma(\mathrm{W_2}(\mathrm{W_1}(\mathrm{GAP(F)})) + \mathrm{W_2}(\mathrm{W_1}(\mathrm{GMP(F)})))\end{aligned} \tag{4}$$

In equation (4): σ represents the activation function Sigmoid, W_1 and W_2 denote the weights of the two fully connected layers, which are then processed by the Relu activation function after W_1, and GAP(F) and GMP(F) represent the global average pooling operation and the global maximum pooling operation, respectively.

2) Spatial Global Attention: The Spatial Global Attention Mechanism, illustrated in Figure 5, is designed to derive spatial attention weights (Ms) by learning the weights of each spatial location. The detailed processing flow is as follows:

Initially, a global average pooling (GAP) operation is applied to the feature map (F) of size W × H × C, generate feature of size W × H × 1, as cross-channel average pooling consolidates the channel count to 1.

Subsequently, effective spatial features are extracted by passing through a convolutional layer with two 1×1 sized convolutional kernels, ensuring the feature map size remains constant.

Ultimately, a two-dimensional spatial attention weight (Ms) is obtained. The spatial domain-based attention module maintains a consistent channel count of 1 in each position, eliminating interference from multiple channels on feature extraction. This ensures the extraction of spatial attention weight information for different positions.

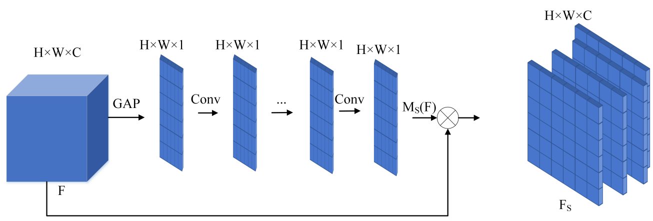

Fig. 5: Spatial Attention Mechanism Module Framework

These spatial attention weights are employed on the feature map, and expression formula for resulting spatial attention feature map F_S delineated in equation (5).

$$\mathrm{F_S} = \mathrm{F} \times \mathrm{M_S}(F) = F \times \sigma(f_2(f_1(GAP(F)))) \tag{5}$$

σ in equation (5) is the activation function Sigmoid, which serves to normalize the values in the [0,1] interval, and both denote convolution operations of size 1 × 1 after each convolution followed by Batch Normalization operation and activation function, and GAP(F) represents the global average pooling operation applied to the feature map.

D. Union loss function

Next stage of the job after Person Re-Identification feature extraction is similarity measure, which measures the distance between the features to evaluate the degree of similarity between two pedestrian images or features. And then determine whether they belong to the same person. Frequently employed metrics to measure loss include Contrast loss and Triple loss. Circle loss, an improved loss function based on Triplet loss, presented at the CVPR 2020 conference, has a more precise convergence objective and a more flexible optimization method is also proposed. In equation (6), L_{circle} for Circle Loss, a_n^j,a_p^i are non-negative weighting factors; s_n is interclass similarity; s_p is intraclass similarity; K represents the quantity of intraclass similarity scores; L is amount of interclass similarity scores; r is the scaling parameter; n and p are acronyms for negative and positive, respectively.

$$\begin{aligned}\mathrm{L_{circle}} &= \log(1 + \sum_{i=1}^{K}\sum_{j=1}^{L}\exp(r(a_n^j s_n^j - a_p^i s_p^i))) \\ &= \log(1 + \sum_{j=1}^{L}\exp(r a_n^j s_n^j)\sum_{i=1}^{K}\exp(-a_p^i s_p^i))\end{aligned} \tag{6}$$

Equation (7) is the cross entropy loss function, where N denotes the amount of pictures in a batch; y^n is the correct label corresponding to each image, and $\hat{y}_i^n$ refers to the predicted label based on the pedestrian features Q_i extracted from each image, which shown in equation (8).

$$\mathrm{L_{ce}}{=}-\sum_{n=1}^{N}\sum_{i} y^n \log(\widehat{y}_i^n) \tag{7}$$

$$\widehat{\mathrm{y}}_i^n = \arg\max_{\mathrm{Z}\in H} \frac{\exp((w_i^z)^T Q_i)}{\sum_{\mathrm{h}=1}^{H}\exp((w_i^h)^T Q_i)} \tag{8}$$

In equation (8): H denotes the total number of label types; z is a category in the sample; w_i^h is a classifier for whether feature Q_i belongs to label h. It is composed of a fully connected layer.

Experts have introduced new method, referred to as PK batch, aimed at enhancing sample efficiency. In this approach, P classes (person IDs) are randomly selected for each batch, followed by the random selection of K images (persons) within each chosen class. Consequently, a total of PK images are amalgamated to create a single batch. Rigorous experiments have demonstrated the significant influence of a meticulously designed triplet loss on the results. This enhanced triplet loss, commonly known as Batch Hard, is defined by the following formula:

$$L_{\mathrm{triplet}} = \sum_{k=1}^{N_K}\sum_{m=1}^{N_M}\Bigg[a + \max_{n=1,\cdots,M} \parallel Q_{k,m}^A - Q_{k,n}^P\|_2 - \min_{\substack{l=1,\cdots,K \\ n=1,\cdots,N \\ l\neq k}} \parallel Q_{k,m}^A - Q_{l,n}^N\|_2\Bigg] \tag{9}$$

In equation (9), N_K is the kind of label in each batch; N_M is the amount of pictures of each type of label in each batch. They have this relationship: N=N_K×N_M; α is a preset value that controls the distance between pairs of positive and negative samples in the feature space; $Q_{i,j}^A$, $Q_{i,j}^P$, $Q_{i,j}^N$ represent the features extracted from the target image, images of the same category, and images of different categories, respectively, where i and j correspond to the label of the image and the index of the image.

To improve the overall network's robustness, this paper suggests new union loss function that combines Ternary Loss Function, Cross Entropy Loss Function, and Circle Loss. Incorporating it at various stages of model training brings the results of similarity measures closer for identical samples and increases the dissimilarity between different samples. The specific formula is:

$$\mathrm{L_{sum}}{=}\omega\mathrm{L_{circle}}{+}\beta\mathrm{L_{b\_triplet}}{+}\lambda(\mathrm{L_{r\_triplet}}{+}\mu\mathrm{L_{ce}}) \tag{10}$$

In equation (10): L_{sum} is the union loss function; $L_{b\_triplet}$ is the triple loss function of the backbone network; $L_{r\_triplet}$ is the triple loss function for relational networks; ω, β, λ and μ are equilibrium parameters.

IV. EXPERIMENTS

A. Experimental Dataset and Evaluation Indicators

1) Experimental Dataset: The datasets selected for the experiments in this study are the Market1501 dataset and DukeMTMC-reID dataset [23]. These dataset details are shown in Table I. The Market1501 dataset is a significant Person Re-Identification dataset obtained from six distinct cameras located across the Tsinghua University campus. In the training set, there were a sum of 12,936 images of pedestrians, involving 751 individuals, with an average of 17.22 training images per person. In the test set, there were 19,732 pictures in total, encompassing 750 pedestrians, with an average of 26.3 test pictures per person. A single image from each camera is randomly chosen to form the Query set, the Query set totaling 3368 images. The DukeMTMC-reID dataset was taken in Duke University and includes a total of 36,411 pictures captured by eight high-resolution cameras. The training set in this dataset contains a total of 16,522 images for 702 identities, with an average of 23.5 training images per individual. The test set contains 2,228 query images from 702 individuals, along with 17,661 images in the pedestrian image library.

TABLE I: Experimental Dataset

DataSet	People	Camera	Pictures	Evaluation
Market1501	1501	6	32217	CMC+mAP
DukeMTMC	1812	8	36441	CMC+mAP

2) Evaluation Index: Person Re-identification is a sorting task that involves matching pedestrian query image features with database image features and then sorting the database images based on similarity from highest to lowest. In this field, the performance of models is frequently assessed using Cumulative Match Characteristic Curve (CMC) and mean Average Precision (mAP).

CMC curves play a crucial role in the field of pattern recognition, serving as a visualization tool for assessing the likelihood that pedestrians in the Person Re-Identification return list match those in the query sample. The CMC curve is essentially the set of cumulative matching rate Rank-n. Assuming a total of N pedestrians, N queries and sorting are performed, and all the query results are denoted as a=(a_1,a_2,a_3,...), the CMC curve can be written as the following equation (11). When i ≤ Rank, m_i = 1; otherwise m_i = 0.

$$\mathrm{CMC}(Rank) = \frac{1}{N}\sum\nolimits_{i=1}^{N} m_i \tag{11}$$

The horizontal coordinate in the CMC plot is Rank, in intervals of positive integers between 1 and n, and the vertical coordinate is the recognition rate Identification of the n images within the recognition results that are most forward in the presence of accurate results. Rank is a case that reflects a value of k in the search results, and for k=1, the curve represents the first hit rate (Rank-1). The CMC curve is the set of results for Rank-k when k takes on different values.

Mean Average Precision Mean (mAP) Unlike Rank-n which only considers the first hit probability, it will consider the full sample and represents the extent to which the person retrieving has all the correct images in the gallery images at the front of the results queue. The mAP curve offers a thorough and impartial evaluation of the model's performance. The sample contains a total of N photos. The count of positive samples with correct prediction is TP (True Positive), the number of positive samples with incorrect prediction is FP (False Positive), the number of negative samples with correct prediction is TN (True Negative), and the number of negative samples with incorrect prediction is FN (False Negative).

Precision indicates the probability that a validated sample has a correct sample and is calculated as in equation (12):

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{12}$$

Average Precision (AP) represents the sum of all accuracies divided by the number of images in the category and measures the model effect on a single category, which is given by equation (13), where C signifies the count of pedestrian IDs and NC signifies the count of images associated with those IDs.

$$AP = \frac{\sum \Pr ecision_c}{N_c} \tag{13}$$

The mAP evaluates the effectiveness of the model overall categories by summing and then averaging the average accuracy of the multi-categorization task, which is given by equation (14):

$$mAP = \frac{\sum_{\mathrm{k=0}}^{\mathrm{c}} \mathrm{AP}_k}{C} \tag{14}$$

B. Experimental Results

The experiment utilized the NVIDIA GeForce GTX 3060 as the hardware platform and operated in the CUDA 11.3 environment. The software platform is built upon the PyTorch 1.12.0 deep learning framework, functioning on the Windows 10 and executed through the PyCharm IDE. In this study, enhancements were made to the ResNet50 algorithm, and experiments were conducted on both the Market1501 dataset and the DukeMTMC-reID dataset, comparing the results with other network models. The results of the experiments are presented in Table II.

Comparative experimental findings indicate that the proposed model in this article boosts the mAP and rank-1 metrics of the proposed algorithm in this paper by 2.4% and 2.9%, respectively, on the Market1501 dataset with HOReID. It also improves by 2.0% and 1.6%, respectively, compared to the more advanced BoT algorithm.

Simultaneously, to illustrate the impacts of multiple models, CMC curves are plotted for the four model structures. Figures 6 and 7 represent pictures of CMC curves on the Market1501 Dataset and the DukeMTMC-reID dataset, respectively.

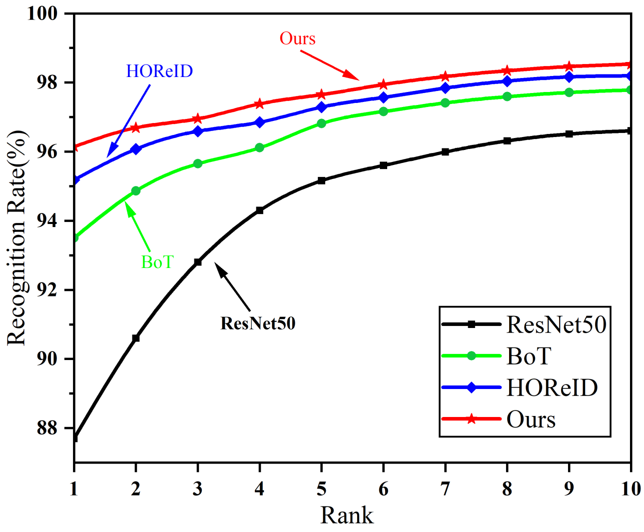

Fig. 6: Market1501 Dataset CMC Plot

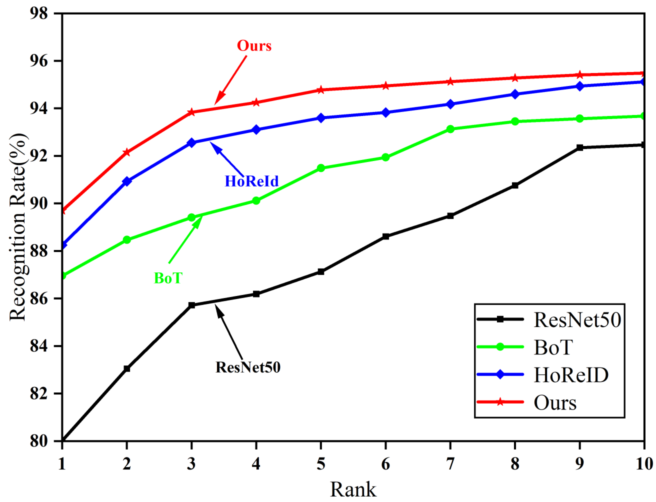

Fig. 7: DukeMTMC-reID Dataset CMC Plot

Based on the findings from Figure 6 and Figure 7, it suggests that the method proposed in this study exhibits varying degrees of improvement on both datasets when compared to ResNet50, BoT, and HOReID, with the most pronounced enhancement observed from Rank-1 to Rank-6. These findings suggest that the enhanced model is capable of extracting more discriminative pedestrian information, effectively overcoming background interference and enhancing the model's adaptability under conditions of scale, occlusion, and attitude changes. Furthermore, it is observed that the overall improvement in the CMC curve of the proposed approach is more prominent on the DukeMTMC-reID dataset than on the Market1501 dataset. This disparity is attributed to the distinct distribution of the two datasets. Despite having a similar total number of images, the DukeMTMC-reID dataset encompasses approximately 20% more pedestrian categories.

Subsequently, the loss rate comparison experiments were performed on the Market1501 dataset for 300 iterations with one taken every five epochs as the results (Iter). The results of the loss rate comparison between ResNet50, PCB algorithm,

TABLE II: Experimental Results of Different Algorithms

Methods	Market1501		DukeMTMC	
	mAP	Rank-1	mAP	Rank-1
ResNet50	71.1	87.8	63.9	80.1
PCB	77.4	92.3	66.1	81.8
PGFA [24]	76.8	91.2	65.5	82.6
HOReID [25]	85.9	93.1	75.6	86.9
OAMN [26]	86.3	93.2	72.6	86.3
BoT	86.3	94.4	77.0	87.2
Ours	**88.3**	**96.0**	**79.8**	**89.8**

and the method proposed in this manuscript in the experiment are shown in Figure 8.

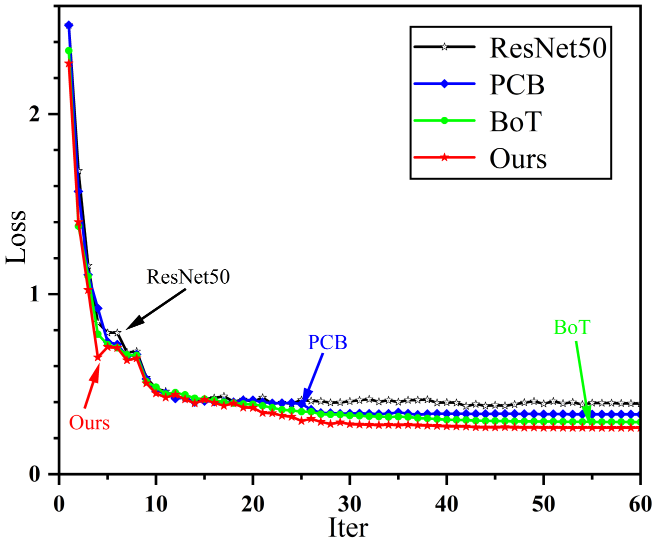

Fig. 8: Loss Ratio Comparison

Figure 8 shows that the Loss value shows a decreasing trend with iterations and stabilizes at 135 iterations, and there is no oscillation in the improved Loss curve. This indicates the viability of the proposed improvement method in this study.

An image is selected as the target image to be retrieved to validate the improved model in this manuscript, and the retrieval results are depicted in Figure 9. In the figure, the depicted query represents the target pedestrian for retrieval. Firstly, extract features from the image library. Subsequently, a similarity measure is applied, and the top 10 results are sorted in descending order of similarity.

As seen in Fig. 9, all retrieval outcomes are correct. And the presence of pedestrians in the figure can be effectively recognized by this paper's algorithm for changes in posture, changes in the surrounding environment, and carrying body accessories.

An ablation experiment is a controlled variable study conducted to assess the effectiveness of proposed methods when multiple enhancements are introduced simultaneously to improve a specific model. The specific outcomes of the ablation experiment conducted in this article are presented in Table III.

The experimental conditions were kept consistent except for the addition or subtraction of modules during the ablation

Fig. 9: Visualization of experimental results

TABLE III: Comparison Table of Ablation Experiment Results

Methods	MarKet1501		DukeMTMC	
	mAP	Rank-1	mAP	Rank-1
R50	71.1	87.8	63.9	80.1
R50+RE	73.5	88.5	66.4	83.6
R50+CA	84.9	92.7	70.3	86.3
R50+UL	78.2	89.2	66.3	85.4
R50+RE+CA	87.9	95.4	77.5	88.3
R50+RE+UL	85.6	94.6	75.2	86.9
R50+CA+UL	88.1	95.9	78.7	**89.8**
R50+All	**88.3**	**96.0**	**79.8**	**89.8**

process. Where R50 represents ResNet50, RE stands for random erasure module, CA stands for Composite Attention Mechanism module, and UL stands for Union loss function. By comparing the models before and after the improvement, the experimental data highlight the validity and value of the new approach.

V. CONCLUSION

In this study, ResNet50 was chosen as the baseline model to improve the pedestrian re-recognition task from three perspectives. Firstly, this article uses a random erasure-based data enhancement technique in the preprocessing stage to generate more data samples. Training the model with more data enhances the model's robustness and generalization. Secondly, a Composite Attention mechanism is added after each residual layer in the network. The separation of background and foreground is accomplished in the spatial domain while filtering interfering features and enhancing the ability to acquire critical channels. Finally, the Triple loss function, cross-entropy loss function, and Circle loss are combined as a joint loss function in the similarity measure stage. The improved model achieves 88.3% mAP and 96.0% Rank-1 on the ResNet50 dataset, which is a notable enhancement over recent models. Experiments demonstrate that the enhanced method proposed in this research can significantly upgrade the performance of the Person Re-identification task. Looking ahead, we will strive to decrease model parameters while not affecting accuracy, so that the model achieves a balance between precision and velocity.

REFERENCES

[1] H. LUO, W. JIANG, and X. FAN, "Research progress on pedestrian re-recognition based on deep learning," *Acta Automatica Sinica*, vol. 45, no. 11, pp. 2032–2049, 2019.

[2] M. Zhu, S. Gong, Z. Qian, S. Serikawa, and L. Zhang, "Person re-identification in the real scene based on the deep learning," *Artificial Life and Robotics*, vol. 26, pp. 396–403, 2021.

[3] H. Gao and Y. Tian, "Research on road-sign detection algorithms based on depth network." *Engineering Letters*, vol. 31, no. 1, pp. 136–142, 2023.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," pp. 1–9, 2015.

[5] C. Chen, B. Wu, and H. Zhang, "An image recognition technology based on deformable and cbam convolution resnet50." *IAENG International Journal of Computer Science*, vol. 50, no. 1, pp. 274–281, 2023.

[6] Q. Cai and J. K. Aggarwal, "Tracking human motion using multiple cameras," vol. 3, pp. 68–72, 1996.

[7] W. Zajdel, Z. Zivkovic, and B. J. Krose, "Keeping track of humans: Have i seen this person before?" pp. 2081–2086, 2005.

[8] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," vol. 2, pp. 1528–1535, 2006.

[9] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," pp. 262–275, 2008.

[10] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," pp. 868–884, 2016.

[11] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, pp. 1–37, 2013.

[12] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," pp. 152–159, 2014.

[13] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," pp. 34–39, 2014.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[15] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov, "Comparison of regularization methods for imagenet classification with deep convolutional neural networks," *Aasri Procedia*, vol. 6, pp. 89–94, 2014.

[16] Y. Liu, Y. Zhang, B. Bhanu, S. Coleman, and D. Kerr, "Multi-level cross-view consistent feature learning for person re-identification," *Neurocomputing*, vol. 435, pp. 1–14, 2021.

[17] G. Chen, T. Gu, J. Lu, J.-A. Bao, and J. Zhou, "Person re-identification via attention pyramid," *IEEE Transactions on Image Processing*, vol. 30, pp. 7663–7676, 2021.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," pp. 7132–7141, 2018.

[19] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," pp. 11 534–11 542, 2020.

[20] W. Chen, Y. Lu, H. Ma, Q. Chen, X. Wu, and P. Wu, "Self-attention mechanism in person re-identification models," *Multimedia Tools and Applications*, pp. 1–19, 2022.

[21] R. Hou, H. Chang, B. Ma, R. Huang, and S. Shan, "Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification," pp. 2014–2023, 2021.

[22] X. Ning, K. Gong, W. Li, and L. Zhang, "Jwsaa: joint weak saliency and attention aware for person re-identification," *Neurocomputing*, vol. 453, pp. 801–811, 2021.

[23] A. Zahra, N. Perwaiz, M. Shahzad, and M. M. Fraz, "Person re-identification: A retrospective on domain specific open challenges and future trends," *Pattern Recognition*, p. 109669, 2023.

[24] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," pp. 4754–4763, 2022.

[25] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," pp. 6449–6458, 2020.

[26] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q. Chen, and R. Ji, "Occlude them all: Occlusion-aware attention network for occluded person re-id," pp. 11 833–11 842, 2021.