ISSN: 2993-8643

Engineering: Open Access

Research Article

A Study of Person Re-Identification Approach Utilizing an Enhanced Convnext Architecture

Xiaoming Sun, [a] Yan Duan, [a] Yan Chen, [a] Junkai Zhang, [a] Yongliang Wang, [a] and Bochao,SU [b*]

[1]*Harbin University of Science and Technology, Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin, China*

[2]*Institute of Intelligent Manufacturing Technology, Shenzhen Polytechnic University, Shenzhen, China*

[*]**Corresponding Author**
Bochao,SU, Institute of Intelligent Manufacturing Technology, Shenzhen Polytechnic University, Shenzhen, China.

Submitted: 2024, Jan 08; **Accepted:** 2024, Feb 12; **Published:** 2024, Mar 01

Citation: Sun, X., Duan, Y., Chen, Y., Zhang, J., Wang, Y. et al. (2024). A Study of Person Re-Identification Approach Utilizing an Enhanced Convnext Architecture. *Eng OA, 2*(2), 01-20.

Abstract

In this paper, we present a new person re-identification method based on an improved ConvNeXt network, called ConvNeXt-AP. This method effectively captures pedestrian features, reduces computing resources, and improves recognition accuracy. ConvNeXt is used as the main network to enhance the capture of local spatial features, with the final forward head module removed to retain more pedestrian features. Moreover, the segmentation strategy from the PCB_RPP network is incorporated at the end of the model to extract fine-grained information from the pedestrian image. To improve the effectiveness of feed-forward convolutional neural networks, the ConvNeXt model Block implements the non-parametric attention mechanism SimAM. This mechanism infers the three-dimensional attention weight of the feature map without adding parameters to the original network, resulting in significant performance improvement with minimal maintenance overhead. To achieve stability and faster convergence, the model achieves stability over time by using a warm-up strategy, and during training, a random erasing strategy is used to reduce the risk of overfitting and improve robustness to occluded pedestrians. Our method has been rigorously tested on multiple public datasets of person re-identification, and the results demonstrate superior performance compared to many state-of-the-art methods.

Keywords: Person Re-Identification; Attention Mechanism; Segmentation Strategy; Warm-Up Strategy; Random Erasing

1. Introduction

Person re-identification, also known as person re-identification, referred to as Reid, is a technology that uses computer vision technology to determine whether there is a specific person in an image or video sequence; in other words, person re-identification refers to the identification of target pedestrians in video sequences that may have been sourced from non-overlapping camera views. It is widely considered as a sub-problem of image retrieval. In addition, person re-identification technology can also be applied to criminal investigation, danger warning, unmanned supermarkets, lost rescue, and traffic monitoring. The traditional pedestrian re-identification method is to manually extract image features, such as color, HOG, SIFT, etc. Then, XQDA or KISSME is used to learn the best similarity measure [1-4]. However, the traditional manual feature description ability is limited, and it is difficult to adapt to the large amount of data tasks in complex scenes. Moreover, in the case of large amount of data, the traditional metric learning method will become very difficult to solve.

In recent years, deep learning represented by convolutional neural networks has achieved great success in the field of computer vision. It has defeated traditional methods in many tasks and even surpassed the human level to some extent [5, 6]. Recently, the Swin Transformer network model proposed by Zeet al [7]. based on the transformer model recently surpasses the CNN model in all areas of computer vision. The model can simulate visual entities of different scales and has a linear computational complexity, while the traditional convolutional network does not have such characteristics; the Swin Transformer can more effectively capture the local features in the image, while the traditional convolutional network can only capture the global features; swim Transformer can better handle visual entities of different scales, while traditional convolutional networks can only handle small-scale visual entities. However, the ConvNeXt network model proposed recently by Zhuang et al [8]. based on the Swin Transformer network model is comparable to the transformer model and performs better in accuracy and scalability. Compared with ViT, the ConvNeXt network is simpler and easier to train and deploy than ViT; the ConvNeXt network is more efficient and can achieve higher performance with less computing resources [9]. the ConvNeXt network can better capture local spatial features, to better handle computer vision tasks, while retaining accuracy, easy deployment, and scalability. The representation learning model based on supervised learning is still a hot topic in current research.

However, in the current state of person re-identification technology, a considerable amount of issues still exist, including diverse feature extraction scales, detecting targets in low light environments, object occlusion, background interference, and target overlap. These issues can adversely affect the accuracy of person recognition. For advancing the person re-identification technology and augmenting the recognition accuracy of pedestrian features while reducing computational resources, this study puts forward an approach of person re-identification that employs the ConvNeXt-AP network, an improved version of the ConvNeXt network that exhibits exceptional performance, simplicity, and efficiency. The principal contributions of this research are the following.

1. Initially, we employ the ConvNeXt network as the backbone network to enhance the localization of spatial features. Subsequently, the final head module of the ConvNeXt network is eliminated to preserve more pedestrian features. Additionally, we incorporate the segmentation strategy from the PCB network at the end of the model to extract fine-grained details from the pedestrian image.
2. The representation capability of feed-forward convolutional neural networks can be substantially enhanced through the inclusion of the non-parametric attention mechanism SimAM in the ConvNeXt model Block. The incorporation of this mechanism enables the inference of the three-dimensional attention weight of the feature map without adding any parameters to the original network. Moreover, it results in significant performance improvement with low maintenance overhead [10].
3. In order to achieve better convergence and overall effectiveness of the network, a warm-up strategy is implemented. This strategy enables the model to gradually stabilize, thus resulting in quicker and more effective convergence. Moreover, during training, the network model incorporates a Random erasing strategy, which proves to be a useful approach that effectively minimizes the risk of overfitting, bolsters the model's resistance to occluded pedestrians and ultimately enhances the model's overall effectiveness [11].

Upon comparison with the state-of-the-art methods, we observe that the approach presented in this study has yielded encouraging outcomes.

2. Related Work

In recent years, person re-identification (Re-ID) has gained significant research interest due to its numerous applications and real-world significance. One of the most actively researched areas in this field is supervised machine learning, which enables full utilization of labeled data and aids in the extraction of discriminative feature representations that result in higher accuracy in person re-identification. Supervised person re-identification models can be broadly divided into three categories: representation learning, metric learning, and ranking learning. Metric learning research endeavors to design effective metric loss functions that improve model performance, while ranking optimization research aims to optimize sequence of results to enhance accuracy. Overall, the central focus in supervised learning is to improve the accuracy of the model by designing effective features and optimizing the performance of its components.

Representation learning is a crucial task in computer vision and can be categorized into four different types: global feature learning, local feature learning, auxiliary feature learning, and video feature learning [12]. Global feature learning involves extracting features from the entire body image. Typically, common approaches of improving global feature learning include leveraging attention mechanisms such as channel attention, spatial attention, as well as cross-image attention, and multi-scale fusion [13-21]. Local feature learning is a method that involves extracting features from localized image regions, which are subsequently combined with whole-body features to produce the final pedestrian features. Various techniques have been developed for this purpose, including the component-based convolution baseline for horizontal partitioning (PCB), pose-driven deep convolution (PDC), multi-scale context-aware network (MSCAN), long-term and short-term memory recursive network, second-order non-local attention, interaction-and-aggregation (IA), the combination of global and local features (GLAD, SCPNet), the adaptive selection of body parts for feature extraction using masks (MaskReID), modeling high-order relationships and topological structures of local features (HOReID), and feature pyramids at multiple scales (HPM) [22-33]. Other introduced methods include randomly dropping some feature map blocks (BDB), weighting the importance of different body parts (CAMA) and adaptively learning features across multiple scales (OSNet) to improve the performance of the algorithms [34-36]. The process of auxiliary feature learning involves incorporating additional information to improve subsequent feature learning stages. This additional information may include pedestrian attributes as semantic information, different orientation within the image as perspective information, representation of data under different cameras as domain information, GAN-generated images as well as random erasing and normalization methods [37-51]. Video feature learning on the other hand utilizes temporally distributed image frames to extract temporal features and combine frames to construct a pedestrian feature expression [52-54].

Metric learning seeks to establish the similarity between two images through the network, while simultaneously mapping the acquired features to a new space [55]. Thus, the aim is to ensure that the distance between two images depicting the same pedestrian (positive sample pairs) is as minimal as possible, and the distance between two images of different pedestrians (negative sample pairs) is as significant as possible. Prior to the emergence of deep learning, metric learning was extensively researched using methods such as the Mahalanobis distance function and the construction of a projection matrix [56-58]. Currently, metric learning utilizes loss function to modify network parameters and enhance image recognition [59]. The prevalent loss functions in current research include Verification Loss, Identity Loss, Triplet loss, and Circle loss [60-66].

Identity Loss treats Re-ID training as an image classification problem, where various images depicting the same pedestrian are categorized into a single group. The common method employed is the SoftMax cross-entropy loss function. Verification Loss considers Re-ID training as an image matching problem, where the objective is to identify whether two images belong to the same pedestrian for binary classification learning. The most commonly used methods are contrasting loss function and binary classification loss function [67, 68]. Triplet Loss framework considers Re-ID training as an image retrieval problem. When comparing same pedestrian images against those of different pedestrians, there exists a greater similarity in feature distance. Circle loss, a novel loss function based on Triplet, aims to address this discrepancy. The primary improvement of Circle loss is that it modifies the original Triplet loss function, which is optimized by the mean force exerted on both positive and negative samples. However, this approach causes difficulty distinguishing between positive and negative samples once the model reaches convergence. On the other hand, Circle loss assigns differing weights to positive and negative samples, which allows for better discrimination between the two. It also regulates the gradient contributions of positive and negative samples, producing a more discriminative model.

Image sorting optimization primarily entails optimizing the order of images retrieved. In general, after extracting the features of a target image and calculating the distance between this target image and the images in the match (Gallery), a ranking result of matched images is obtained [69-70]. This ranking result is used to calculate the Rank-n Accuracy and draw the CMC curve. This process may include re-ranking and rank fusion techniques. Re-ranking involves optimizing the initial sorting list by utilizing inter-library similarity. Sorting fusion combines multiple ranking lists generated from disparate approaches to improve retrieval performance [71-73]. Our goal is to enhance pedestrian feature capture while reducing computational resources and improving person re-identification accuracy. To accomplish this, we analyze the advantages and disadvantages of supervised learning, and introduce a new pedestrian re-identification method based on an improved ConvNeXt network, named ConvNeXt-AP network.

3. Approach

In this section, we present the flowchart in Fig. 1 which illustrates the proposed person ReID model. The training phase comprises four modules: preprocessing, training strategy, network model, and classification. Image preprocessing is required to reduce interferences in input features. The completion of training set transformation enriches the set, improves the model's generalization ability, and reduces interference feature introduction while enhancing model robustness. The neural network is highly unstable at the start of training. To mitigate risk of overfitting, two training strategies are employed, namely, learning rate preheating and random erasing. These strategies enhance the model's robustness and promote convergence to occlusion. The network model module employs an improved ConvNeXt network called ConvNeXt-AP network. This network model removes the last ConvNeXt network head module, retaining more pedestrian features. Additionally, the segmentation strategy in the PCB_RPP network is incorporated at the model's end to extract fine-grained pedestrian image information. The addition of the non-parametric attention mechanism, SimAM, to the ConvNeXt model block significantly improves the ability of the network to extract pedestrian image features while incurring low overhead. Pedestrian re-identification is accomplished through the classification module.

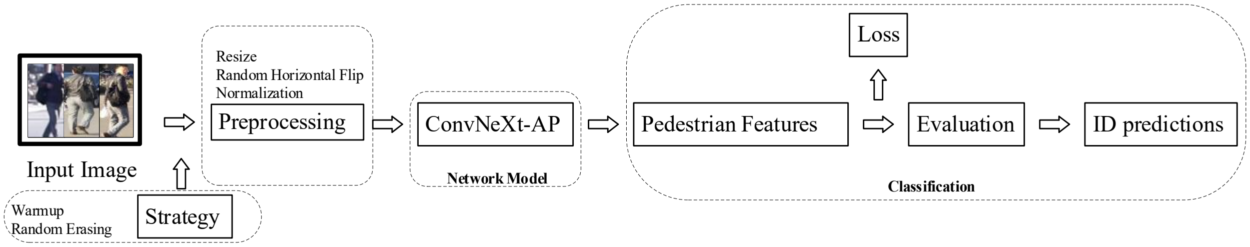


Figure 1: Person ReID flowchart.

3.1 Strategy

3.1.1 Warm-up

The learning rate warm-up strategy refers to gradually increasing the learning rate during the initial stages of training a neural network model. This strategy is employed to address the network's instability at the beginning of training, aiming to facilitate better convergence. By starting with a smaller learning rate and gradually increasing it, the warm-up strategy helps the model adapt more effectively to the features of the training data, thereby improving training efficiency and performance. An example of the learning rate warm-up process is illustrated in Fig.2.The learning rate warm-up not only aids in model convergence and prevents getting stuck in local optima but also enhances the speed of convergence and overall performance.

As shown in Figure.2., in this example, the total number of epochs is set to 50, warm-up epochs is set to 5, lr_init is set to 0.02, and lr_max is set to 0.1. In the learning rate warm-up strategy, lr_init represents the initial learning rate during the warm-up phase, gradually increasing the learning rate for faster model convergence. On the other hand, lr_max is a relatively large maximum learning rate used after the warm-up period to continue training the model for improved results. During the first 5 epochs (the warm-up phase), the learning rate gradually increases from 0.02 to 0.1, speeding up the model's convergence. After the warm-up phase, the learning rate gradually decreases to maintain training stability. Additionally, in the later stages of the example, the model employs a cosine annealing strategy to adjust the learning rate. This strategy gradually reduces the learning rate in a cosine function manner as the training progresses. Unlike other learning rate decay strategies, the cosine annealing ensures a smoother decrease in the learning rate, preventing the model from failing to converge due to a rapid decline in the learning rate.

The main purpose of the learning rate warm-up strategy is to accelerate convergence and improve training efficiency in the initial stages of training. The cosine annealing strategy, implemented in the later stages, helps prevent oscillations and overfitting by gradually adjusting the learning rate. In summary, employing these two strategies during model training effectively enhances stability, training speed, convergence, and generalization capabilities of deep learning models.

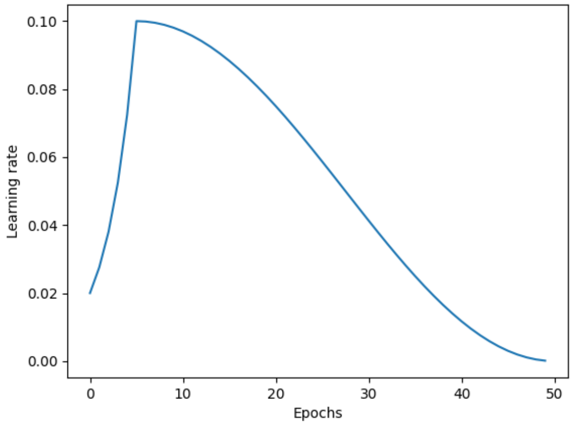


Figure 2: Learning rate change using Cosine Warm-up.

3.1.2 Random Erasing

The random erasure strategy is a technique to augment data for training deep learning models. During the training process, this method randomly selects a rectangular area in the image and erases its pixels with random values, thus generating training images with varying degrees of occlusion to reduce overfitting and increase robustness to occlusion. This is illustrated in Fig. 3, where after random erasing, certain parts of the original image are covered by a rectangular box filled with random values to simulate occlusion and enhance the sample data. By adopting the random erasing strategy, the risk of overfitting can be effectively reduced and the model can be made more robust to occlusions, thereby improving its overall performance.

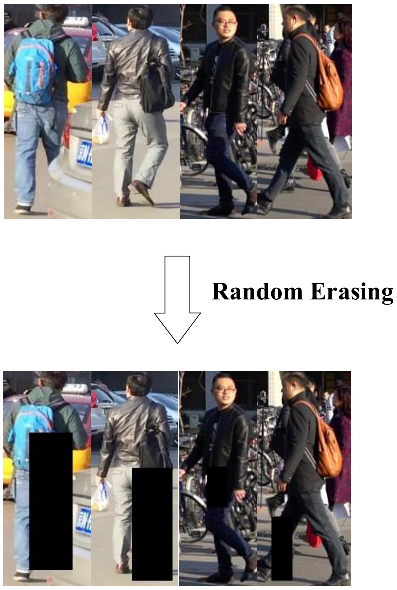


Figure 3: Random Erasing.

3.2 Preprocessing

The multitude of factors, including the large or small image sizes in the dataset, the various pedestrian postures, differing image resolutions, occlusions, and changes in image domain can contribute to interference features during pedestrian feature extraction. These features can result in less effective algorithms. Hence, image preprocessing is necessary to minimize these disturbances. Data augmentation, also known as image preprocessing, enriches the training set with transformed data to improve the model's generalization ability. In this study, we employed the image preprocessing methods of Resize, Random Horizontal Flip, and Normalization during model training to alleviate interference features. The aforementioned techniques provided robustness to the model by reducing the introduction of interference features.

3.3 Network Model：ConvNeXt-AP

In recent years, Transformer-based models, such as ViT, have shown superior performance compared to traditional CNN models in various computer vision tasks. Swin Transformer, a notable example of Transformer-based models, has drawn significant attention due to its remarkable performance. Unlike traditional convolutional networks, Swin Transformer can simulate visual entities at varying scales, while achieving linear computational complexity. Swin Transformer models excel at capturing local features in images, which traditional convolutional networks, designed for capturing global features, struggle with. Additionally, Swin Transformer models demonstrate better performance in handling visual entities of different scales compared to traditional convolutional networks that can only handle small-scale visuals.

Despite the superiority of Transformer-based models over traditional CNN models, Zhuang et al. introduced ConvNeXt, a novel CNN model that draws inspiration from Swin Transformer based on ResNet architecture. ConvNeXt outperforms Swin Transformer in image classification and detection segmentation tasks, while still maintaining similar scalability as vision transformers, thus achieving better performance as data volume and model size increase. ConvNeXt networks present simpler and more efficient alternatives to ViT, as they are easier to train, deploy, and require fewer computing resources while achieving higher accuracy. Unlike vision transformer models, ConvNeXt models excel in capturing local spatial features, making them more suitable for a variety of computer vision tasks. Additionally, ConvNeXt models exhibit high accuracy and scalability, while being easy to deploy, making them ideal candidates for various computer vision applications.

The article proposes a methodology, named ConvNeXt-AP, that aims to improve the recognition accuracy of person re-identification, capture local spatial features, and reduce computing resources requirements. This method relies on an improved ConvNeXt network architecture, presented in Figure. 4, which is simple, efficient and highly effective. The ConvNeXt network architecture is employed in the methodology as the backbone network, to capture local spatial features better. The forward_head module of ConvNeXt is removed to retain more person features. At the end of the model, the segmentation strategy from the PCB network is introduced to extract fine-grained information from the person image. The refined part pooling (RPP) strategy from PCB is utilized to address the problem of including extreme values in the fragments during uniform partitioning. The RPP redistributes these extreme values to more similar fragments to enhance the content consistency of each fragment, leading to better model performance. To enhance the feedforward convolutional neural network's representation ability effectively, the article introduced the SimAM module, a parameter-free attention mechanism, in the ConvNeXt model block, introduced in Fig. 5. Significantly, the SimAM module can infer the three-dimensional attention weight of the feature map without adding parameters to the original network, leading to considerable performance improvements with low costs, resulting in better model performance.

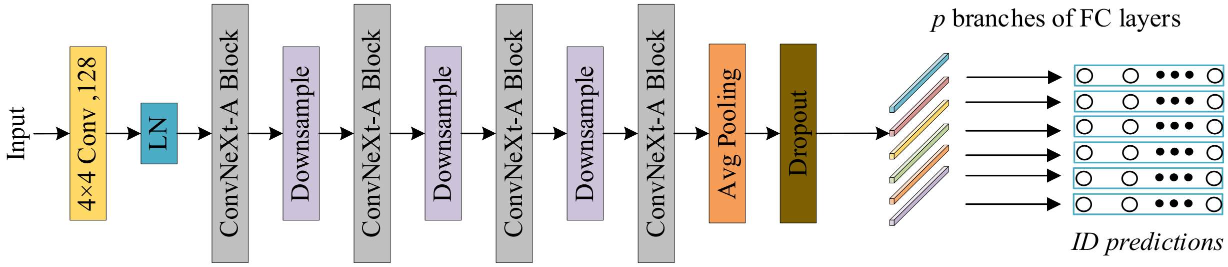


Figure 4: The ConvNeXt-AP Network.

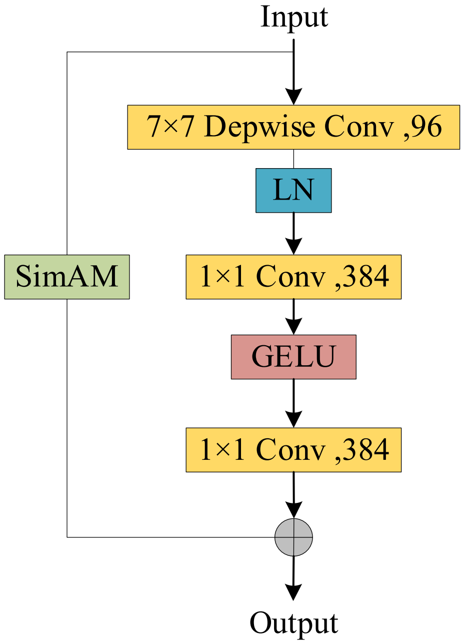


Figure 5: ConvNeXt-A Block.

3.3.1 ConvNeXt Network

To improve the ConvNeXt model's performance in capturing local spatial features while reducing the network's FLOPs and thereby enhancing its efficiency in handling computer vision tasks, the model's macro design is divided into two main parts. Firstly, the stacked block ratio, initially 1:1:2:1 in ResNet50, is adjusted to 1:1:3:1. Secondly, Patchify replaces the conventional downsampling module stem, typically comprising a convolutional layer with a 7x7 kernel size and a stride of 2, followed by a max-pooling layer with a 2 stride - both downsampling both height and width by a factor of 4. Patchify, like Swin Transformer, uses a large, non-overlapping convolutional layer for downsampling. These design modifications enhance the ConvNeXt model's ability to capture local spatial features efficiently while reducing FLOPs and handling computer vision tasks more effectively.

In ConvNeXt, the group convolution employs cross-group interactions. Each group's output interacts with the outputs of other groups to capture global feature information, thus enhancing the representation power of the network. This results in an improved representation power of the network, without an increase in computational complexity. Furthermore, ConvNeXt uses depthwise convolution, which reduces computational load, and improves model accuracy. Although the number of channels could be increased from 64 to 96, ConvNeXt increases it from 32 to 64.

The bottleneck layer structure in ConvNeXt model, initially used in ResNet, consists of a small middle and two large ends aiming to reduce the number of parameters and FLOPs. Conversely, MobileNet v2's structure includes a large middle and two small ends, to prevent information loss. In addition, the MLP module in Transformer is analogous to the inverse bottleneck structure. ConvNeXt also employs the inverse bottleneck layer structure, leading to decreased parameter and FLOPs count, thereby enhancing the network's efficiency.

The use of smaller kernel sizes in mainstream CNN architectures can effectively decrease the number of parameters and computational complexity. Additionally, 3x3 convolutions have GPU efficient implementations. Swin-T, on the other hand, uses a larger window size of 7x7, which can capture more local details than the 3x3 convolution window. ConvNeXt also explores the use of larger kernel sizes. The results of experiments indicate that increasing the kernel size offers potential to improve the model's performance. Nonetheless, a 7x7 convolutional kernel is selected to capture more local details after the performance has saturated following experiments.

At the micro level, ConvNeXt modifies the activation function by replacing ReLU with GELU to align it with other metrics. Transformer and ConvNeXt employ only a few activation functions with Transformer using the activation function solely on one MLP, and ConvNeXt adding GELU activation functions only between two 1x1 convolutions. ConvNeXt derives insight from Transformer and includes Batch Normalization only after the first convolution, swaps out all other instances of BN for Layer Normalization, and introduces a separate downsampling layer. Consequently, these micro designs lead to decreased parameter and FLOPs count, enhance the network's efficiency, improve model convergence, and reduce overfitting.

3.3.2 The Segmentation Strategy

The segmentation strategy can utilize any classification network without a hidden fully connected layer. In order to balance network simplicity and performance, the forward_head module was removed from the end of the ConvNeXt network. The segmentation strategy was added thereafter to extract granular information regarding the pedestrian images. Fig. 6 illustrates this approach.

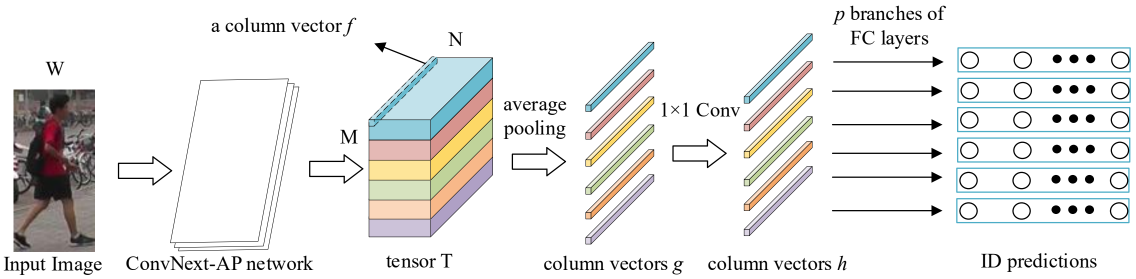


Figure 6: The Segmentation Strategy.

1) Initially, the removed forward_head module of ConvNeXt network is employed as the backbone.
2) The image of size 384×128×3 is processed by the backbone to obtain a tensor T of size 12×6×f along the channel dimension.
3) The tensor T is then partitioned into p horizontal stripes, where an average pooling (AP) operation is performed on f, resulting in g.
4) Convolutional layers are utilized to lower the dimensionality and produce h, which is 256- dimensional.
5) The classifier takes each h and is fine-tuned using a softmax optimization technique to predict the ID. The softmax loss function is applied separately for each classification layer to optimize their weights.

Additionally, we incorporate the refined part pooling (RPP) strategy of the PCB model to tackle the issue of extreme values being present in fragments during standard partitioning. RPP redistributes these extreme values among fragments with similar content to improve the content consistency. As a result, this technique enhances the content consistency of each fragment and, in turn, heightens the model's performance. This is shown in Fig. 7.

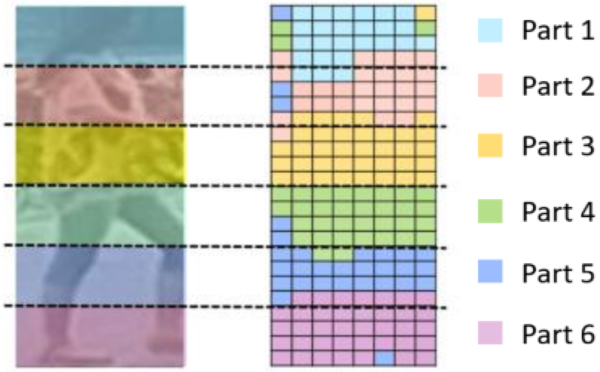


Figure 7: A rendering of RPP strategy.

The RPP strategy reassigns extreme values present in f to similar parts depending on the resemblance between f and gi. The steps involved in this implementation are as follows:

1) part classifier:
Initially, the part classifier computes the probability of f being present in region Pi by using the weight matrix W.

$$P(P_i \mid f) = \mathrm{softmax}\left(W_i^T f\right) = \frac{\exp\left(W_i^T f\right)}{\sum_{j=1}^{p} \exp\left(W_j^T f\right)} \qquad (1)$$

2) sampling operation:
Then, according to the probability P $(P_i \mid f)$ of f belonging to the region P_i, a new region partition is performed with P as the sampling weight of P_i.

$$P_i = \{P(P_i \mid f) \times f, \forall f \in F\} \qquad (2)$$

The PCB+RPP structure is obtained by replacing the avg pooling steps in the PCB structure with the two steps 1) and 2) described above, Where Pi represents the i-th horizontal partition. As shown in Fig. 8.

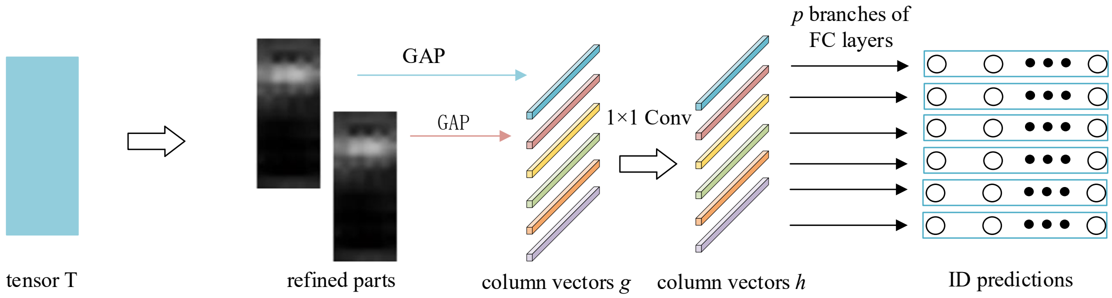


Figure 8: PCB+RPP structure.

3.3.3 The SimAM Attention Module

SimAM is a novel attention module for Convolutional Neural Networks (CNN) that is compatible with existing architectures such as ResNet and VGG, and delivers superior performance in multiple computer vision tasks. SimAM is a simple yet effective approach that infers the three-dimensional attention weights of feature maps without adding parameters to the original network, resulting in no memory overhead or extra computational resources.

We integrate the parameter-free attention mechanism SimAM into the ConvNeXt model block, as illustrated in Figure 5. This effectively enhances the model's performance and robustness while reducing computational costs.

Most of existing attention modules generate 1-D or 2-D weights from features X,and then expand the generated weights for channel (a) and spatial (b) attention. As shown in Figure 9.

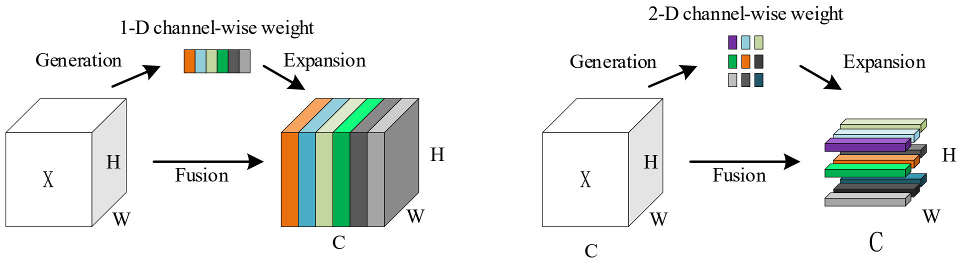


Figure 9 **1**-D or 2-D Attention Module.

Full 3-D weights is better than conventional 1-D and2-D attentions. SimAM Attention Module propose to refine that features with full 3-D weights, as shown in Fig. 10.

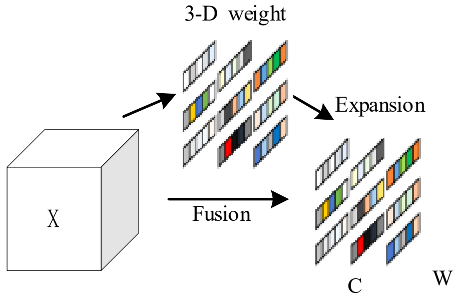


Figure 10: SimAM Attention Module.

The simAM attention module assigns a unique weight to each neuron. The following energy function for each neuron:

$$e_t(w_t, b_t, \mathbf{y}, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}\left(-1-(w_t x_i + b_t)\right)^2 + \left(1-(w_t t + b_t)\right)^2 + \lambda w_t^2 \qquad (3)$$

$$\mathrm{w_t} = \frac{2(\mathrm{t} - \upmu_\mathrm{t})}{(\mathrm{t} - \upmu_\mathrm{t})^2 + 2\sigma_\mathrm{t}^2 + 2\lambda} \tag{4}$$

$$\mathrm{b_t} = -\frac{1}{2}(\mathrm{t} + \upmu_\mathrm{t})\mathrm{w_t} \tag{5}$$

Where w_t and $b_t$are weight and bias the transform .Where t and x_i are the target neuron and other neurons in a single channel of the input feature.X ∊ RC×H×W. i is index over spatial dimension and M = H × W is the number of neurons on that channel. $\upmu_\mathrm{t} = \frac{1}{\mathrm{M}-1}\sum_{\mathrm{i}=1}^{\mathrm{M}-1}\mathrm{x_i}$ and $\sigma_\mathrm{t}^2 = \frac{1}{\mathrm{M}-1}\sum_{\mathrm{i}}^{\mathrm{M}-1}(\mathrm{x_i} - \upmu_\mathrm{t})^2$ are mean and variance calculated over all neurons except t in that channel.

The simAM attention module use a scaling operator rather than an addition for feature refinement. The whole refinement phase of our module is:

$$\tilde{\mathrm{X}} = \mathrm{sigmoid}\left(\frac{1}{\mathrm{E}}\right) \odot \mathrm{X} \tag{6}$$

where E groups all e_t*across channel and spatial dimensions.
In short, it can be easily inserted into convolutional neural networks to improve the capture of image detail features.

3.4 Loss Functions

In supervised learning, regardless of whether it is a regression or a classification problem, employing a loss function is essential. The loss function measures the inconsistency between the predicted value f(x) and the actual value y of the model. If the loss function is small, it implies that the machine learning model is proximate to the real data distribution, and, therefore, performs well. If the loss function is large, it implies that the machine learning model differs significantly from the true data distribution and thereby performs poorly. The main training objective is to apply optimization methods to determine the model parameters that correspond to the loss function's least value. So, in conjunction with the benefits of the Id loss and Triple loss, Circle loss is introduced as an alternative loss function to enhance the accuracy, efficiency, and differentiation of the pedestrian re-identification algorithm. Circle loss is a radius-based loss function utilized mainly for deep neural network classification tasks. Its primary objective is to address the issue of imbalanced intra-class distribution and unclear inter-class distance while also preventing overfitting. Circle loss applies weight to the positive and negative sample pairs to regulate the gradient involvement of each sample pair. Ultimately, an enhanced model with greater discriminatory power can be achieved.

Given K intra-class similarity scores and L inter-class similarity scores, the formula for Circle loss is derived by maximizing intra-class similarity and minimizing inter-class similarity. The formula is as follows:

$$\mathrm{L_{circle}} = \log\left[1 + \sum_{\mathrm{i}=1}^{\mathrm{K}}\sum_{\mathrm{J}=1}^{\mathrm{L}}\exp\left(\gamma\left(\alpha_\mathrm{n}^\mathrm{j}\mathrm{s}_\mathrm{n}^\mathrm{j} - \alpha_\mathrm{p}^\mathrm{i}\mathrm{s}_\mathrm{p}^\mathrm{i}\right)\right)\right] \tag{7}$$

$$\mathrm{L_{circle}} = \log\left[1 + \sum_{\mathrm{i}=1}^{\mathrm{K}}\exp\left(\gamma\alpha_\mathrm{n}^\mathrm{j}\mathrm{s}_\mathrm{n}^\mathrm{j}\right) - \sum_{\mathrm{J}=1}^{\mathrm{L}}\exp\left(-\gamma\alpha_\mathrm{p}^\mathrm{i}\mathrm{s}_\mathrm{p}^\mathrm{i}\right)\right] \tag{8}$$

where γ is a scaling factor, α_n^j 、 α_p^i are non-negative weighting factors, s_n^j is the inter-class similarity, and s_p^i is the maximization of intra-class similarity.

Circle loss has several advantages, including its ability to address the issues of uneven intra-class distribution and unclear inter-class distances. By projecting comparable samples to neighboring locations, Circle loss can increase the margin between classes and enhance the model's classification accuracy. Moreover, Circle loss can also prevent overfitting by preventing the model from overfitting on the training dataset and increasing its generalization capacity. A combination of traditional performance metrics, including Rank-n, CMC, and mAP, are employed to enhance the precision, efficiency, and significance of pedestrian re-identification algorithms.

4 Data and Realization

To validate the efficiency of the ConvNeXt-AP network, we carried out experiments on three widely utilized person ReID datasets.

4.1 CUHK03

The CUHK03 dataset is widely used in person re-identification research and comprises pedestrian images taken on the Chinese University of Hong Kong (CUHK) campus. The dataset was obtained from six stationary surveillance cameras capturing each pedestrian identity from two non-overlapping camera views, resulting in an average of 4.8 images per pedestrian for each camera view. The dataset contains 1467 pedestrians, the training set consisting of 767 entities and the testing set consisting of 700 entities. For each pedestrian in the testing set, one image is picked randomly from all cameras to form the query set, while the other pedestrian images in the testing set comprise the gallery set [74].

4.2 Market-1501

The Market-1501 dataset was collected in 2015 during the summer at Tsinghua University campus, comprising 1501 pedestrians and 32,668 detected pedestrian bounding boxes captured by six cameras, one low-resolution and five high-resolution. At least two cameras captured each pedestrian, and in some cases, there exist multiple images of the same person in one camera. The training dataset is composed of 751 individuals and includes 12,936 images in total, with an average of 17.2 training images per individual. The 750-individual test set includes 19,732 images, with an average of 26.3 test images per individual [75].

4.3 DukeMTMC-reID

The DukeMTMC-reID dataset is a substantial pedestrian image dataset designed for person re-identification, which became available in 2017 [76]. It was primarily obtained from 8 static cameras positioned across Duke University campus. The DukeMTMC-reID dataset has been frequently adopted as a benchmark dataset in the field of reID, with numerous models proposed by both academia and industry. The dataset encompasses 36,411 images of 1,812 unique pedestrians, with 1,404 distinct pedestrians captured by more than two cameras, and 408 pedestrians detected by only one camera. Comparative tables for datasets are provided in Table 1.

Dataset	Pedestrian number	Image number	Annotation method	Camera number	Multi-camera capturing
CUHK03	1,467	13,164	mixed	10	Yes
Market1501	1,501	4,096	mixed	6	Yes
DukeMTMC-reID	1,812	36,441	manual	8	Yes

Table 1: Dataset Comparison.

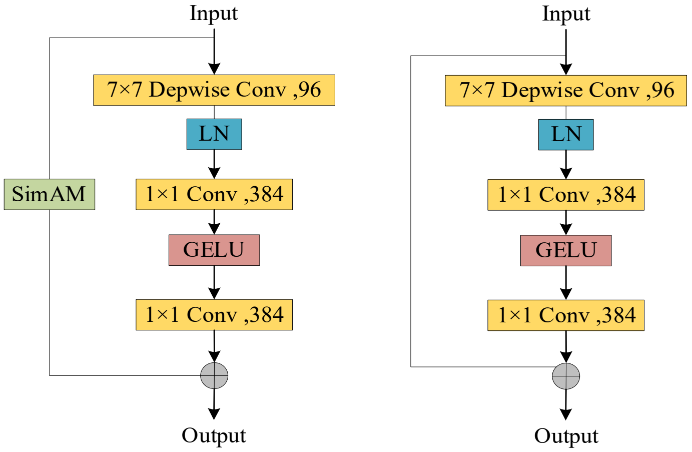
5 Experimental Results

5.1 Implementation Details

The experimentation employed a Windows 11 operating system, with PyTorch 1.11 framework, and programmed in Python 3.7. The processing units utilized were an AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz CPU, and an NVIDIA GeForce RTX 3090 with 24GB of memory GPU. A preprocessing strategy was implemented to eliminate interfering features during pedestrian feature extraction, which extensively hampers the algorithm's efficacy. The preprocessing strategy includes Random Erasing, Resize, Random Horizontal Flip, and Normalization. Random Erasing's p parameter is set to 0.5, and Resizing modifies the pedestrian image dimensions into h, w = 256, 128. The paper also introduced a learning rate warm-up strategy to ensure that the network can converge well. The warm-up strategy's epochs were set to 120, the warm-up_epochs were set to 5, with lr_init and lr_max set to 0.02 and 0.1, respectively.

5.2 Analysis of The Attention Module

The experiment conducted an incremental evaluation and validation to ascertain the effects of Attention Modules for ConvNeXt-AP on three datasets; CUHK03, Market-1501, and DukeMTMC-reID. The following two variants of the Block were constructed: (a) Block and (b) Block + simAM. Fig. 11 displays the comparison between the two variants.



a) ConvNeXt-A Block.

b) ConvNeXt Block

Figure 11: The comparison between the two Block.

Based on Table 2, when Attention Modules are absent, the baseline model has a rank-1 accuracy of merely 72.5%, 92.9%, and 83.9%, with corresponding mAP scores of 70.8%, 81.7%, and 73.4% on the CUHK03, Market1501, and DukeMTMC-reID datasets, respectively. Implementing Attention Module enhances the model's outcome, leading to a rank-1 improvement of 0.8%, 0.6%, and 1.4%, with an enhancement in mAP of 1.4%, 0.9%, and 1.8%, respectively.

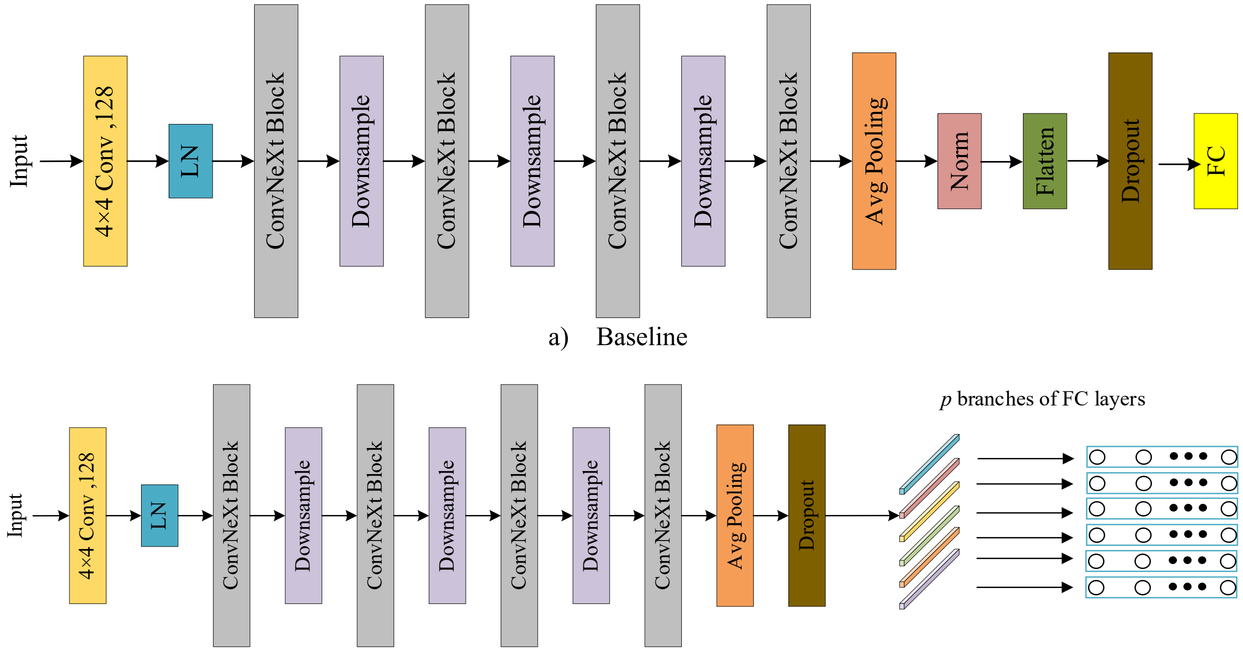
	CUHK03		Market-1501		DukeMTMC-reID			
Model	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	FLOPs	Parameters
Baseline	72.5	70.8	92.9	81.7	83.9	73.4	10.04G	88.00 M
Baseline + simAM	73.3	72.2	93.5	82.6	85.3	75.2	10.04 G	88.00 M

Table 2: Attention Module Ablation Experiment.

SimAM, an innovative convolutional neural network attention module, determines the 3D attention weights of feature maps without adding parameters to the original network. This process helps decrease computational expenses, accelerate model speed and promote the model's precision, robustness, and performance to handle noise and perturbations in the image. Comparative analyses of the FLOPs and parameters show that the difference between the two models is minimal, further supporting the above conclusion. In summary, SimAM facilitates easy integration into convolutional neural networks, leading to improved image detail feature extraction.

5.3 Analysis of The Segmentation Strategy

The paper performed incremental evaluation and validation on three datasets - CUHK03, Market- 1501, and DukeMTMC-reID to determine the effects of the Segmentation Strategy on ConvNeXt-AP. Two variants were constructed based on the baseline model - (a) Baseline, and (b) Baseline + the Segmentation Strategy. Fig. 12 presents a comparison between the two variants.



a) Baseline

b) Baseline + the Segmentation Strategy

Figure 12: The comparison between the two Baseline.

The results presented in Table 3 indicates that the baseline achieves a rank-1 accuracy of only 72.5%, 92.9%, and 83.9%, with corresponding mAP scores of 70.8%, 81.7%, and 73.4% on the CUHK03, Market1501, and DukeMTMC-reID datasets, respectively, when the Segmentation Strategy is not used. Implementing the Segmentation Strategy enhances the model's outcome, leading to rank-1 accuracy improvements of 3.3%, 1.8%, and 3.3%, with an enhancement in mAP of 2.8%, 3.4%, and 2.5%, respectively.

	CUHK03		Market-1501		DukeMTMC-reID			
Model	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	FLOPs	Parameters
Baseline	72.5	70.8	92.9	81.7	83.9	73.4	10.04 G	88.00 M
Baseline+ Strategy	75.8	73.6	94.7	85.1	87.2	75.9	10.04 G	90.74 M

Table 3: Attention Module Ablation Experiment.

The segmentation strategy is primarily used for the extraction of features. It involves dividing the image into various parts during the feature extraction process and concatenating the derived features to derive a comprehensive feature map. Part-based Convolutional Baseline (PCB) divides the input image into several parts and extracts features from each of these parts. Residual Pyramid Pooling (RPP) is a feature pooling method used to enhance the representational power of the feature maps. RPP applies pyramid pooling on the feature map and merges the pooled outcomes with the initial feature map using concatenation. Comparing the FLOPs and parameters, we notice little difference in FLOPs between the two models mentioned above, whereas the improved model has more parameters. Nevertheless, the research findings affirm the previous statement. The segmentation strategy improves the models' overall performance by making them more robust to factors such as pedestrian posture and occlusion. Additionally, it helps improve the representational power of the feature map, making it an efficient technique for pedestrian re-identification.

5.4 Analysis of The ConvNeXt-AP Network

To evaluate the effects of the ConvNeXt-AP network, we conducted incremental validation and evaluation on three datasets: CUHK03, Market-1501, and DukeMTMC-reID. We constructed three variants of the baseline model: "Baseline (ConvNeXt)," "Baseline + Attention Module + Segmentation Strategy (ConvNeXt-A)," "Baseline + Attention Module (ConvNeXt-P)," and "Baseline + Attention Module + Segmentation Strategy (ConvNeXt-AP)."

	CUHK03		Market-1501		DukeMTMC-reID			
Model	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	FLOPs	Parameters
The ConvNeXt	72.5	70.8	92.9	81.7	83.9	73.4	10.04G	88.00 M
The ConvNeXt-A	73.3	72.2	93.5	82.6	85.3	75.2	10.04 G	88.00 M
The ConvNeXt-P	75.8	73.6	94.7	85.1	87.2	75.9	10.04 G	90.74 M
The ConvNeXt- AP	76.9	74.8	95.2	86.5	88.2	77.3	10.04 G	90.74 M

Table 4: Ablation Experiment.

Table 4 shows that without the ConvNeXt-AP network and the two strategies, the baseline only achieves rank-1 scores of 72.5%, 92.9%, and 83.9% on the CUHK03, Market1501, and DukeMTMC-reID datasets, respectively. Adding these two strategies improves rank-1 by 4.4%, 2.3%, and 4.3% for the respective datasets. Similarly, it improves mAP by 4.0%, 4.8%, and 3.9%. Based on the results presented above, integrating the two strategies into the ConvNeXt network offers the following benefits:

1. Improved recognition accuracy: The partitioning strategy decomposes the image into different components for recognition purposes, and the attention mechanism facilitates a more accurate comparison of similarities between these components, resulting in more accurate recognition of various pedestrians within the image.

2. Increased robustness: Combining the two strategies reduces the risk of overfitting and enhances the overall robustness of the model.

In summary, combining the partitioning strategy and the attention mechanism strategy and incorporating them into the ConvNeXt network can achieve better recognition performance, stronger robustness, and faster training speed.

5.5 Comparison with State-of-the-Art Methods

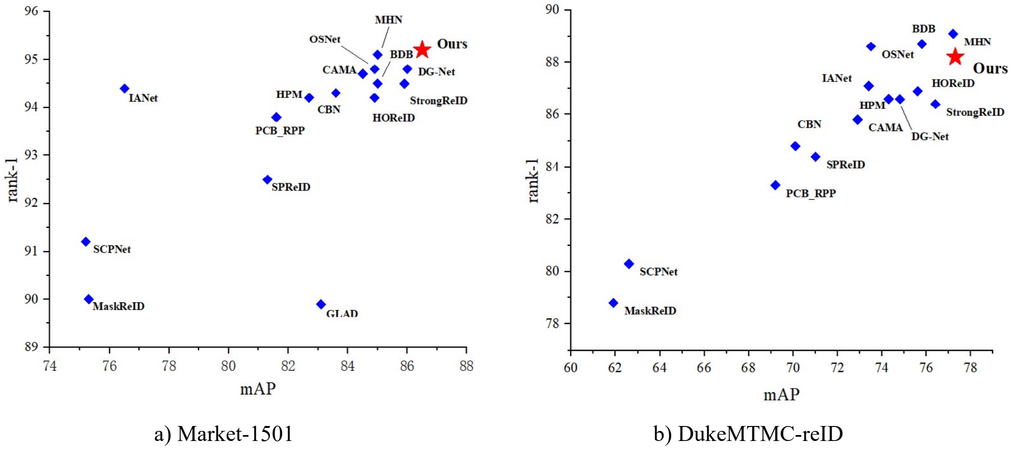
To demonstrate the advanced pedestrian re-identification performance of the ConvNeXt-AP network, we compared it with state-of-the-art methods on two datasets: Market-1501, and DukeMTMC-reID, as shown in Table 5.

	Market-1501		DukeMTMC-reID	
Method	Rank-1	mAP	Rank-1	mAP
GLAD[29]	89.9	83.1	—	—
MaskReID[31]	90	75.3	78.8	61.9
SCPNet[30]	91.2	75.2	80.3	62.6
SPReID[41]	92.5	81.3	84.4	71
PCB-RPP[22-23]	93.8	81.6	83.3	69.2
HOReID[32]	94.2	84.9	86.9	75.6
HPM[33]	94.2	82.7	86.6	74.3
CBN[50]	94.3	83.6	84.8	70.1
IANet[28]	94.4	76.5	87.1	73.4
StrongReID[51]	94.5	85.9	86.4	76.4
BDB[34]	94.5	85	88.7	75.8
CAMA[35]	94.7	84.5	85.8	72.9
DG-Net[48]	94.8	86	86.6	74.8
OSNet[36]	94.8	84.9	88.6	73.5
MHN[20]	95.1	85	**89.1**	77.2
Ours	**95.2**	**86.5**	88.2	**77.3**

Table 5: Comparison of results on Market1501, and DukeMTMC-reID.

The results demonstrate the high performance of the ConvNeXt-AP network on pedestrian re-identification tasks by comparing it to other state-of-the-art methods on the Market-1501 and DukeMTMC-reID datasets, including models based on global and local features, mask-based models, semantic segmentation-based models, models based on interaction and aggregation, pose-based models, generative-based models, and attention-based models. Table 5 shows that the ConvNeXt-AP network outperforms most other methods in both Rank-1 and mAP, particularly on the DukeMTMC-reID dataset. The above conclusion is more intuitively demonstrated in Figure 13.

In particular, on the Market-1501 dataset, the ConvNeXt-AP network achieved a Rank-1 accuracy of 95.2% and an mAP of 86.5%. For Market-1501, the Rank-1 accuracy improved by 0.1% and mAP improved by 1.5% compared to the MHN method. On the DukeMTMC-reID dataset, ConvNeXt-AP achieved a Rank-1 accuracy of 88.2% and an mAP of 77.3%, which is one of the best results among all existing methods. Although its Rank-1 accuracy is not the best on the DukeMTMC-reID dataset, ConvNeXt-AP outperforms all other methods in terms of mAP.



a) Market-1501

b) DukeMTMC-reID

Figure 13: Comparison with State-of-the-Art Methods on different datasets

By achieving the best results on both datasets, this study demonstrates the ConvNeXt-AP network's outstanding performance in pedestrian re-identification tasks. Furthermore, the ConvNeXt-AP network exhibits superior performance compared to existing methods, indicating its feasibility and practicality. These results provide valuable references for future research.

5.5 Experimental Visualization

In this section, we present experimental results by visualizing probe images and identifying the top five matching gallery images. We conducted experiments using the ConvNeXt-AP network on the Market1501 and DukeMTMC-reID datasets. The correct matches are highlighted with a green bounding box in Figure 14-15. Our proposed model can accurately retrieve corresponding pedestrian images from the gallery, even with occluded, profile, or back-view probe images.

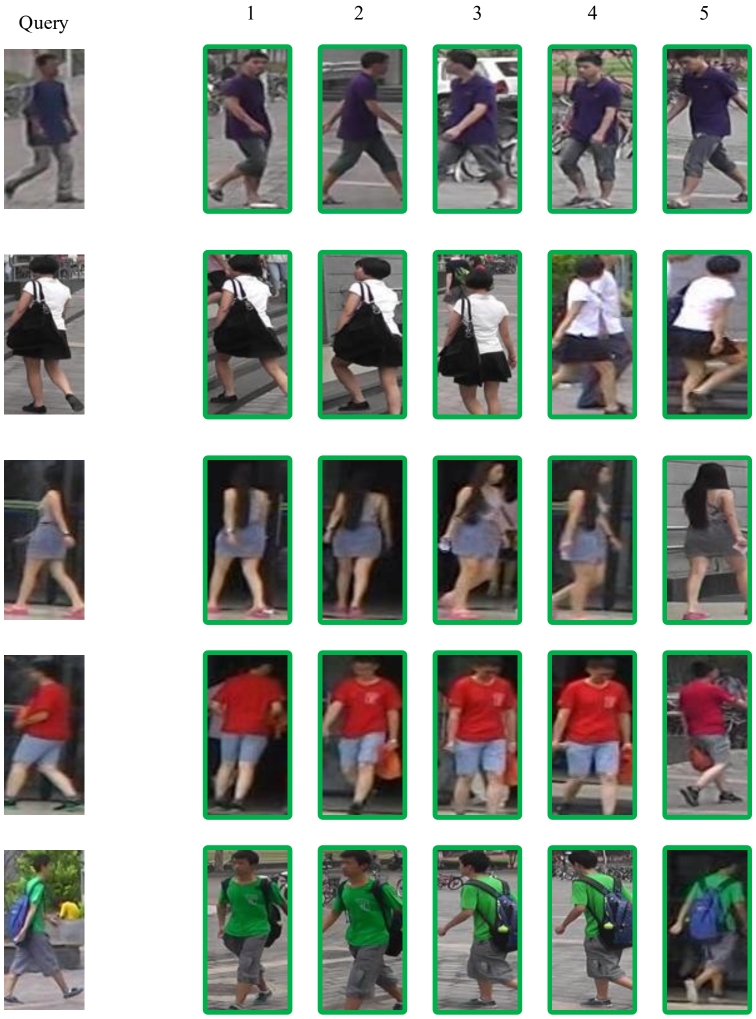


Figure 14: The top five rankings on query images of the Market1501 dataset using the proposed ConvNeXt-AP network.

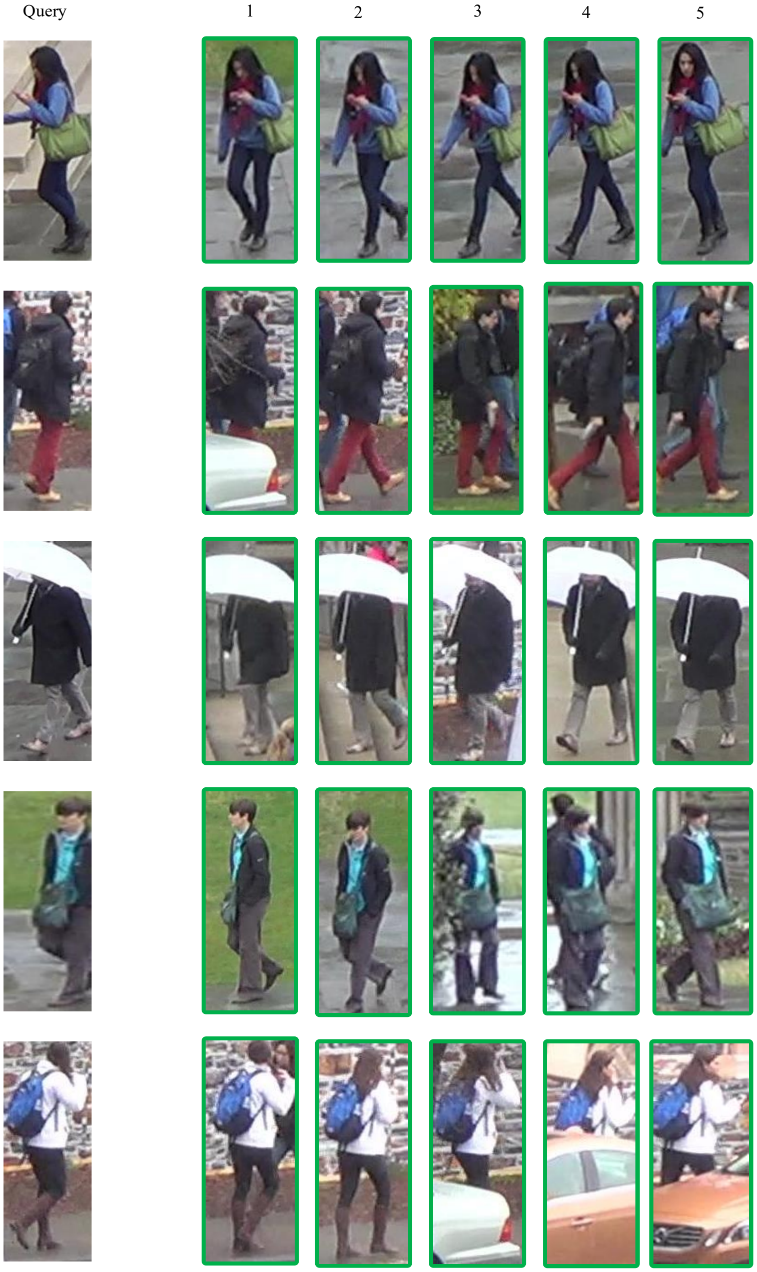


Figure 15: Pedestrian feature heatmaps.

In order to visualize the attention regions and important features of ConvNeXt-AP network on pedestrian images more intuitively, we employed feature heatmaps and compared them with the traditional ConvNeXt network, as shown in Figure 16. By generating feature heatmaps on pedestrian images, we can visually observe the differences in attention levels of ConvNeXt-AP network and ConvNeXt network towards different body parts of pedestrians. The ConvNeXt-AP network improves the attention capability towards different body parts in pedestrian images by introducing the slice strategy and parameter-free attention module. Compared to the traditional ConvNeXt network, the ConvNeXt-AP network more accurately captures the feature information in pedestrian images. The generated feature heatmaps reveal how the ConvNeXt-AP network focuses on the key features of different body parts in pedestrian images, demonstrating the advantages of ConvNeXt-AP network in pedestrian re-identification tasks.

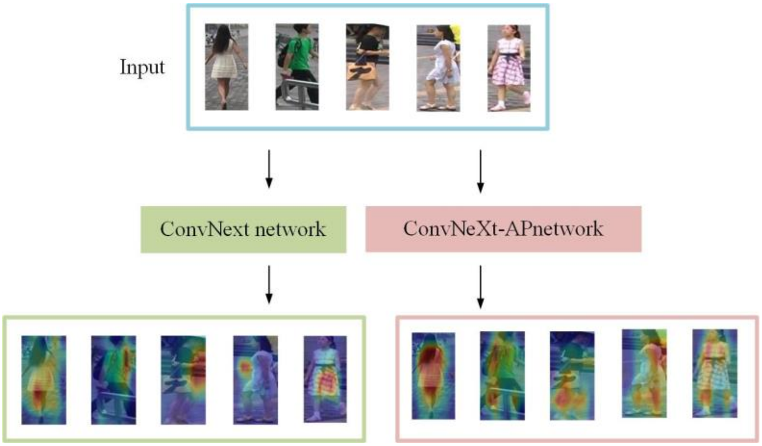
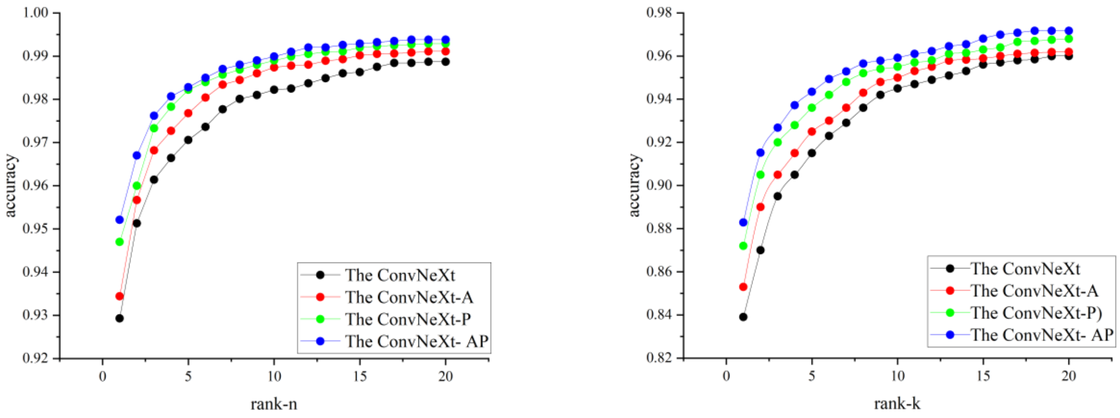


Figure 16: Pedestrian feature heatmaps.

Additionally, we plot the cumulative matching characteristic (CMC) curves in Figure 17 to present our model's performance at different rank values. The CMC curve is a commonly used metric for evaluating model performance at different rank values.

In summary, our experimental visualization comprehensively demonstrates the high model performance in pedestrian re-identification task.



(a)Market-1501

(b)DukeMTMC-reID

Figure 17: The CMC curve graphs on different datasets: (a) Market-1501 and (b) DukeMTMC-reID.

In conclusion, the superiority of the ConvNeXt-AP network in pedestrian re-identification tasks has been thoroughly demonstrated through the visualization experiments, including matching pedestrian visualizations, feature heatmaps, and CMC curve visualizations.

6. Conclusions

This paper proposes a pedestrian re-identification method called ConvNeXt-AP, which is based on an improved ConvNeXt network. Our approach aims to capture pedestrian-related features more effectively while improving re-identification accuracy. The backbone network is implemented using the ConvNeXt network, which can better capture local spatial features. Additionally, we remove the forward_head module to retain more pedestrian-related features. At the end of the model, we introduce a segmentation strategy to extract fine-grained information from pedestrian images. This strategy increases the model's robustness to factors such as pedestrian pose and occlusion, and it also improves the representational capacity of feature maps, thus enhancing the model's performance. It makes our approach an effective pedestrian re-identification method.

To improve the representational capacity of convolutional neural networks (CNNs) effectively, the SimAM, a parameter-free attention mechanism, is incorporated into the ConvNeXt model block. SimAM can infer three-dimensional attention weights of a feature map without adding parameters to the original network, resulting in a significant boost in accuracy and robustness to noise and disturbances while keeping computation time low. This method reduces computational complexity, improves model speed, and ultimately enhances the model's performance and accuracy in capturing image detail features. In summary, the SimAM can be easily added to CNNs to improve the representation of images.

In order to ensure successful network convergence, a learning rate warm-up strategy was introduced. This gradually stabilizes the model, enhancing convergence speed and optimizing performance results. Additionally, the training process utilizes a random erasing strategy which reduces overfitting risks and equips the model with greater resilience against occluded pedestrians, thus ensuring higher performance outcomes.

Our study has conducted experiments on various datasets and analyzed the results. Based on our analysis, we have reached the following conclusions:

1. The results of our study suggest that utilizing a segmentation strategy to divide image inputs into multiple parts, combined with an attention mechanism that draws similarity-based comparisons between such parts, can significantly enhance the accuracy rate of pedestrian recognition.

2. Our study has found that incorporating both the learning rate warming strategy and the random erasing strategy into the model training process can lead to a stronger level of robustness. This is due to their ability to minimize the risk of overfitting, thus improving the model's generalization to novel data.

3.Our study has demonstrated that utilizing a combination of four specific strategies - segmentation, attention, learning rate warm-up, and random erasing - can make the training process more efficient for pedestrian recognition models. Specifically, this combination allows the model to better leverage the training data, leading to better performance in fewer iterations and a reduction in overall training time.

In conclusion, we have proposed a model that combines a segmentation strategy, attention mechanism strategy, random erasing strategy, and learning rate warm-up strategy with the ConvNeXt network that significantly exceeds the performance of current state-of-the-art models.

Acknowledgments

This study is funded by the Scientific Research Startup Fund for Shenzhen High-Caliber Personnel of Shenzhen Polytechnic (No. 6023330002K), the Guangdong Provincial Key R&D Programme (No. 2020B0909020002), the Guangdong Provincial General University Innovation Team Project (No. 2020KCXTD047), the National Natural Science Foundation of China (No. U1913206), and the College Projects of Shenzhen Polytechnic (No. 6021320035K and 6023310026K). Supported by General Higher Education Project of Guangdong Provincial Education Department, No.2023KCXTD077 Supported by college start-up fund of ShenZen PolyTechnic University,No.6022312031K

References

1. Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886-893). Ieee.
2. Lowe, D. G. (1999, September). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision* (Vol. 2, pp. 1150-1157). Ieee.
3. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012, June). Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2288-2295). IEEE.
4. Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2197-2206).
5. Gao, P., Yue, X., Chen, W., Fang, W., Tian, Z., & Zhang, F. (2022). A state-of-the-art review on person re-identification with deep learning. *International Journal of Ad Hoc and Ubiquitous Computing, 41*(2), 69-91.
6. Liu, M., Zhao, J., Zhou, Y., Zhu, H., Yao, R., & Chen, Y. (2022). Survey for person re-identification based on coarse-to-fine feature learning. *Multimedia Tools and Applications, 81*(15), 21939-21973.
7. Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976-11986).
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.*
9. Yang, L., Zhang, R. Y., Li, L., & Xie, X. (2021, July). Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning* (pp. 11863-11874). PMLR.
10. Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020, April). Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 13001-13008).
11. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., & Tian, Q. (2017). Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1367-1376).
12. Li, W., Zhu, X., & Gong, S. (2018). Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2285-2294).
13. Wang, C., Zhang, Q., Huang, C., Liu, W., & Wang, X. (2018). Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 365-381).
14. Shen, Y., Xiao, T., Li, H., Yi, S., & Wang, X. (2018). End-to-end deep kronecker-product matching for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6886-6895).
15. Wang, Y., Chen, Z., Wu, F., & Wang, G. (2018). Person

re-identification with cascaded pairwise convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1470-1478).

16. Chen, G., Lin, C., Ren, L., Lu, J., & Zhou, J. (2019). Self-critical attention learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9637-9646).
17. Chen, D., Xu, D., Li, H., Sebe, N., & Wang, X. (2018). Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8649-8658).
18. Luo, C., Chen, Y., Wang, N., & Zhang, Z. (2019). Spectral feature transformation for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4976-4985).
19. Chen, B., Deng, W., & Hu, J. (2019). Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 371-381).
20. Qian, X., Fu, Y., Jiang, Y. G., Xiang, T., & Xue, X. (2017). Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 5399-5408).
21. Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)* (pp. 480-496).
22. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., ... & Sun, J. (2017). Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184.*
23. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2017). Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 3960-3969).
24. Li, D., Chen, X., Zhang, Z., & Huang, K. (2017). Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 384-393).
25. Varior, R. R., Shuai, B., Lu, J., Xu, D., & Wang, G. (2016). A siamese long short-term memory architecture for human re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14* (pp. 135-153). Springer International Publishing.
26. Xia, B. N., Gong, Y., Zhang, Y., & Poellabauer, C. (2019). Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3760-3769).
27. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019). Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9317-9326).
28. Wei, L., Zhang, S., Yao, H., Gao, W., & Tian, Q. (2017, October). Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 420-428).
29. Fan, X., Luo, H., Zhang, X., He, L., Zhang, C., & Jiang, W. (2019). Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14 (pp. 19-34). Springer International Publishing.
30. Qi, L., Huo, J., Wang, L., Shi, Y., & Gao, Y. (2019, July). A mask based deep ranking neural network for person retrieval. In *2019 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 496-501). IEEE.
31. Wang, G. A., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., ... & Sun, J. (2020). High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6449-6458).
32. Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., ... & Huang, T. (2019, July). Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 8295-8302).
33. Dai, Z., Chen, M., Gu, X., Zhu, S., & Tan, P. (2019). Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3691-3701).
34. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., & Zhang, S. (2019). Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1389-1398).
35. Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3702-3712).
36. Su, C., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2016). Deep attributes driven multi-camera person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14* (pp. 475-491). Springer International Publishing.
37. C.-P. Tay , S. Roy , and K.-H. Yap, “Aanet: Attribute attention network for person re-identifications,” in CVPR, (2019).
38. Zhao, Y., Shen, X., Jin, Z., Lu, H., & Hua, X. S. (2019). Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4913-4922).
39. Wang, J., Zhu, X., Gong, S., & Li, W. (2018). Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2275-2284).
40. Kalayeh, M. M., Basaran, E., Gökmen, M., Kamasak, M.

E., & Shah, M. (2018). Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1062-1071).

41. Chang, X., Hospedales, T. M., & Xiang, T. (2018). Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2109-2118).
42. Liu, F., & Zhang, L. (2019). View confusion feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6639-6648).
43. Zhu, Z., Jiang, X., Zheng, F., Guo, X., Huang, F., Sun, X., & Zheng, W. (2020, April). Aware loss with angular regularization for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 13114-13121).
44. Lin, J., Ren, L., Lu, J., Feng, J., & Zhou, J. (2017). Consistent-aware deep learning for person re-identification in a camera network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5771-5780).
45. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., & Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 994-1003).
46. Li, Y. J., Lin, C. S., Lin, Y. B., & Wang, Y. C. F. (2019). Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7919-7929).
47. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., & Kautz, J. (2019). Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2138-2147).
48. Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020, April). Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 13001-13008).
49. Zhuang, Z., Wei, L., Xie, L., Zhang, T., Zhang, H., Wu, H., ... & Tian, Q. (2020). Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16* (pp. 140-157). Springer International Publishing.
50. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., & Gu, J. (2019). A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia, 22*(10), 2597-2609.
51. McLaughlin, N., Del Rincon, J. M., & Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1325-1334).
52. Zhou, Z., Huang, Y., Wang, W., Wang, L., & Tan, T. (2017). See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4747-4756).
53. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., & Zhou, P. (2017). Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 4733-4742).
54. Kaya, M., & Bilge, H. Ş. (2019). Deep metric learning: A survey. *Symmetry, 11*(9), 1066.
55. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012, June). Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2288-2295). IEEE.
56. Zheng, W. S., Gong, S., & Xiang, T. (2011, June). Person re-identification by probabilistic relative distance comparison. In *CVPR 2011* (pp. 649-656). IEEE.
57. Liao, S., & Li, S. Z. (2015). Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE international conference on computer vision* (pp. 3685-3693).
58. Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014, August). Deep metric learning for person re-identification. In *2014 22nd international conference on pattern recognition* (pp. 34-39). IEEE.
59. Varior, R. R., Shuai, B., Lu, J., Xu, D., & Wang, G. (2016). A siamese long short-term memory architecture for human re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14* (pp. 135-153). Springer International Publishing.
60. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., & Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 994-1003).
61. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., & Tian, Q. (2017). Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1367-1376).
62. Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J., ... & Wang, X. (2018). Eliminating background-bias for robust person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5794-5803).
63. Guo, Y., & Cheung, N. M. (2018). Efficient and deep person re-identification using multi-level similarity. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2335-2344).
64. Ye, M., Lan, X., Wang, Z., & Yuen, P. C. (2019). Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security, 15,* 407-419.
65. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., & Wei, Y. (2020). Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6398-6407).
66. Varior, R. R., Shuai, B., Lu, J., Xu, D., & Wang, G. (2016).

A siamese long short-term memory architecture for human re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14* (pp. 135-153). Springer International Publishing.

67. Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 152-159).
68. Ye, M., Liang, C., Wang, Z., Leng, Q., & Chen, J. (2015, October). Ranking optimization for person re-identification via similarity and dissimilarity. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1239-1242).
69. Liu, C., Loy, C. C., Gong, S., & Wang, G. (2013). Pop: Person re-identification post-rank optimisation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 441-448).
70. Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1318-1327).
71. McLaughlin, N., Del Rincon, J. M., & Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1325-1334).
72. Ye, M., Liang, C., Yu, Y., Wang, Z., Leng, Q., Xiao, C., ... & Hu, R. (2016). Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia, 18*(12), 2553-2566.
73. Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 152-159).
74. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision* (pp. 1116-1124).
75. Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016, October). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision* (pp. 17-35). Cham: Springer International Publishing.