Mathematical Biosciences and Engineering

MBE, 20 (2): 2110–2130.
DOI: 10.3934/mbe.2023098
Received: 04 September 2022
Revised: 25 October 2022
Accepted: 07 November 2022
Published: 14 November 2022

http://www.aimspress.com/journal/MBE

Research article

TwinsReID: Person re-identification based on twins transformer's multi-level features

Keying Jin[1], Jiahao Zhai[2] and Yunyuan Gao[2,*]

[1] Zhuoyue Honors College, Hangzhou Dianzi University, Hangzhou, China
[2] College of Automation, Hangzhou Dianzi University, Hangzhou, China

*** Correspondence:** Email: gyy@hdu.edu.cn; Tel: +13757187567.

Abstract: In the traditional person re-identification model, the CNN network is usually used for feature extraction. When converting the feature map into a feature vector, a large number of convolution operations are used to reduce the size of the feature map. In CNN, since the receptive field of the latter layer is obtained by convolution operation on the feature map of the previous layer, the size of this local receptive field is limited, and the computational cost is large. For these problems, combined with the self-attention characteristics of Transformer, an end-to-end person re-identification model (twinsReID) is designed that integrates feature information between levels in this article. For Transformer, the output of each layer is the correlation between its previous layer and other elements. This operation is equivalent to the global receptive field because each element needs to calculate the correlation with other elements, and the calculation is simple, so its cost is small. From these perspectives, Transformer has certain advantages over CNN's convolution operation. This paper uses Twins-SVT Transformer to replace the CNN network, combines the features extracted from the two different stages and divides them into two branches. First, convolve the feature map to obtain a fine-grained feature map, perform global adaptive average pooling on the second branch to obtain the feature vector. Then divide the feature map level into two sections, perform global adaptive average pooling on each. These three feature vectors are obtained and sent to the Triplet Loss respectively. After sending the feature vectors to the fully connected layer, the output is input to the Cross-Entropy Loss and Center-Loss. The model is verified On the Market-1501 dataset in the experiments. The mAP/rank1 index reaches 85.4%/93.7%, and reaches 93.6%/94.9% after reranking. The statistics of the parameters show that the parameters of the model are less than those of the traditional CNN model.

Keywords: person re-identification; Transformer; deep learning; multi-level features; self-attention

1. Introduction

Person Re-identification (ReID) is an important research direction in the field of computer vision. It uses computer vision technology to determine whether there is a specific person in an image or video sequence. Its research is aimed at the retrieval and identification of specific target pedestrians. It is automatic target identification and recognition technology, which mainly describes in the surveillance environment of multiple cameras without overlapping fields of view. ReID essentially describes whether an interested target person who has appeared under one lens reappears under another camera through a correlation algorithm.

Common challenges of ReID problems include body dislocation, occlusion, back-ground disturbance, viewing angle changes, posture changes, and noise tags. In the past few years, there have been many excellent ReID models based on CNN. The general process of the implemented model is: first extract the features of the input image from the convolutional neural network, obtain the feature representation of the image after dimensionality reduction, convolution, pooling, etc, and train to optimize the classification loss and metric learning loss on the training data set. Then calculate the Euclidean distance between each picture in the query set to be queried, and get the predicted output of several top-k pictures. All in all, the person re-identification problem can be attributed to the classification problem and the supervised learning of metric learning. At present, the mainstream experimental paradigms mainly include CNN and Transformer in the ReID direction.

CNN has many advantages. For example, it uses convolution kernels to continuously extract high level abstract features. In theory, its receptive field should cover the entire image, but many studies [1] have shown that its actual receptive field is much smaller than the theoretical receptive field. It is not conducive to making full use of context information to capture features. Although it is possible to continuously stack deeper convolutional layers, it will obviously cause the model to be too bloated and the amount of calculation will increase sharply.

The advantage of the Transformer lies in the use of attention to capture global contextual information so as to establish a long-distance dependence on the target, thereby extracting more powerful features. At present, some research [2] have proved that Transformer had a better migration effect, and premodels on large data sets could be well migrated to small data sets. The Transformer model is originally applied to Natural Language Processing (NLP). The network structure abandons the traditional convolutional neural network and recurrent neural network. And it is entirely composed of attention mechanisms. To be precise, Transformer only consists of self-attention and feed forward neural networks. Models are widely used in the NLP field, such as machine translation, question answering systems, text summarization, and speech recognition. In the Transformer model [3,4] in the NLP field, the Transformer is mainly composed of encoder-decoder and self-attention, which can handle the logical relationship between input sentence sequences and get a good predictive output. Since Transformer can only predict the input sequence, it cannot be directly applied to the field of image computer vision. Therefore, the Google team proposed the famous ViT [2] image classification model without changing the Transformer structure. The input image is divided into several patches. These patches are spliced into an image sequence, then a classification vector token is learned in the encoded-decoding layer of the model, and finally training the classification. ViT abandoned the

traditional CNN classification model and used Transformer's self-attention mechanism for learning, providing a new experimental paradigm for the computer vision field. Since then, various variant models based on Transformer have emerged endlessly.

However, it is not suitable to directly apply Transformer to dense prediction tasks. A picture often needs at least a few hundred pixels to express information, while modeling a long sequence data is precisely the inherent flaw of Transformer. Based on the idea of grouped convolution and separable convolution in CNN, MeiTuan team proposed locally grouped self-attention (LSA) and global subsampled attention (GSA) [5]. Among them, LSA captures more detailed short distance information, while GSA captures long distance and global information. Then combined the advantages of Transformer and CNN, a CNN like hierarchical backbone network Twins-SVT Transformer is designed. And it is implemented by Transformer.

Although CNN has been widely used in various engineering fields [6,7], the calculation of CNN still deepens channel information by stacking convolutional layers, resulting in high computational complexity and long model training iteration cycle. The essence of Transformer model is to calculate the correlation between each input to obtain the corresponding prediction result, which has low computational complexity.

In order to solve the problem that traditional CNN networks excessively rely on convolutional operation and cost a lot of computation, this paper takes Twins-SVT Transformer as ReID's backbone, extracts the hierarchical information and fine-grained features in the transformer, and carries out experiments combining with the final features of the backbone. The main content of this paper is to explore the feasibility of transformer replacing CNN in ReID field and the integration of different levels of features to improve the model performance.

The main contributions of this article:

- Apply Twins-SVT as the backbone of ReID and study the feasibility and the effectiveness of the Transformer architecture in visual retrieval tasks.
- Propose the ReID retrieval model twinsReID based on Twin-SVT. The feature maps of two different stages of Twins-SVT Transformer are processed to obtain three feature vectors with different stages and different fine-grained sizes. Metric training and classification training are performed on all the obtained feature vectors to obtain the optimal solution of the model.

2. Related work

Since Transformer made breakthroughs in NLP tasks, researchers in the industry have been trying to use Transformer in the Computer Vision (CV) field. In various down-stream tasks of computer vision, many excellent models based on Transformer have emerged.

In the field of image classification, iGPT [8] and Vision Transformer [2] are all models that use Transformer structure to classify images. ViT retains the NLP Transformer structure to the greatest extent. After cutting the picture into several patches, add relative position coding, and train the classification token after passing through the Transformer's encoder-decoder. Deit [9] is an image Transformer based on knowledge distillation proposed by Facebook Research. The author added a new distillation token, which uses the self-attention layer to interact with the class token and patch token to improve the performance of the model. TNT [10] is a Transformer-based model architecture proposed by Huawei Noah Labs in 2021. TNT jointly extracts the local and global features of the image through two internal and external Transformers and converts the image into patch embedding sequence and

pixel embedding sequence by stacking TNT modules. There is also the realization of the Transformer in the task of object detection. DERT [11] is a Transformer-based object detection model proposed by Facebook Research in 2020. The feature map obtained through CNN is flattened into a one dimensional feature map, and the position information is added to the encoder of Transformer. Several embedding features output by the encoder are decoded into bounding box coordinates in parallel through the action of the decoder. In the field of image semantic segmentation, Transformer performs equally well. Segmenter [12] is based on the research results of ViT, which divides the image into patches, maps them into linear embedded sequences, and encodes them with an encoder. Then Mask Transformer decodes the output of the encoder and class embedding, and finally classifies each pixel after upsampling. In the ReID task, there are also models based on ViT. TransReid [13] is a work of ReID research based on the Transformer structure. SIE (Side Information Embedding) and JPM (Jigsaw Patch Module) are proposed in the research. SIE includes Side information such as viewing angle, camera shooting style, etc. The essence of JPM is to randomly group the embedding vector obtained by each patch, then scramble and divide it. And finally get several local features combined with the output of the standard transformer as global information features for joint training.

To adapt to the task of CNN, many scholars are also studying the backbone model based on Transformer. Swin Transformer [14] is a pyramidal model improved by Microsoft based on Transformer. Similar to the structure of CNN's ResNet [15], the corresponding feature map is obtained through multi-layer network modules. Its core idea is to limit the self-attention calculation to nonoverlapping partial windows, and perform self-attention calculation in the window area without overlap while allowing cross window connection. Twins is the Transformer model proposed by MeiTuan. The author replaced the positional encoding in PVT [16] with the Conditional Positional Encodings proposed by the team in CPVT [17]. The experiment proved that it can be used in classification and downstream tasks. It obtained substantial performance improvement, especially on dense tasks. Since the conditional position encoding (CPE) supports variable-length input, the visual Transformer can flexibly handle features from different spatial scales.

The success of Swin Transformer and Twins proves that Transformer has great potential in the field of vision. It is an alternative to CNN and has expanded application prospects for various computer vision tasks. In the ReID task, the focus is on the processing and optimization of the feature map. This article attempts to build a model based on Twins-SVT. First, input the picture 224×224×3 into Twins-SVT to obtain the feature map of 7×7×512, and then obtain the vector f_g after pooling. Taking into account the Transformer's self-attention mechanism, the rest of the attention mechanism is not added in the feature extraction stage, but the multi level feature representation is considered. First take out the feature map of 14×14×256 of stage3 of Twins-SVT. After convolution operation, get the feature map of 512 channels, split it horizontally, and get two feature vectors f_{l1} and f_{l2} after pooling. Then pass these three vectors through the BN layer, and obtain the predicted output p_g, p_{l1} and p_{l2} from the fully connected layer. Send p_g, p_{l1}, p_{l2} to the Cross entropy loss, and f_g, f_{l1}, f_{l2} are sent to Triplet loss, and then combined with Center loss for training. In the prediction and reasoning period, the three feature vectors are connected as the feature descriptor of the input picture.

3. Model introduction

For the ReID model based on the CNN backbone, its parameters have been at a relatively high level due to the continuous accumulation of network layers and the large number of convolutional

operations. Based on the Transformer architecture, the model has a strong application in the CV field because of its convenient computing process, and self-attention mechanism. The reason why Transformer based models such as ViT, DERT, etc. perform well on various tasks is the Self-Attention mechanism. The Self-Attention mechanism enables the model to learn what it should pay attention to in each input.

3.1. Self-attention

The Attention mechanism [18] was proposed by the Bengio team in 2014 and has been widely used in various fields of deep learning in recent years. For example, it is used in computer vision to capture the receptive field on an image, or it is used in NLP to locate key tokens or features. The Transformer is essentially an Encoder-Decoder structure, which is a conversion model that completely relies on self-attention to calculate the input and output representations without using sequence aligned with recurrent neural networks or convolutional neural networks.

The realization of the core self-attention module in Transformer is to learn Q, K, and V vectors separately for a certain input, and then calculate the Attention. The formula is as shown in formula (1).

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \times V \tag{1}$$

Where d_k is the dimension of the vector, and its function is to scale the vector.

In order to improve the Attention mechanism and enhance the ability to express Attention, Google proposed Multi-Head Attention [18], a multi-head attention mechanism, which can be compared to the simultaneous use of multiple convolution kernels in CNN. Intuitively speaking, multi-head attention mechanism helps the network to capture richer characteristic information. The specific implementation is to linearly transform the Q, K, and V of each head through the parameter matrix, and the parameter matrix of each head is not shared, and then calculate attention, and finally repeat *h* times to splice the results. The formula is shown in formula (2).

$$\begin{cases} head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \\ MultiHeadAtten(Q,K,V) = Concat(head_1, \ldots, head_h)W^O \end{cases} \tag{2}$$

Where $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W_i^O \in \mathbb{R}^{(h \times d_v) \times d_k}$, $d_k = d_v = d_{model/h}$.

3.2. Twins-SVT transformer

At present, Transformer based models are mostly used for image classification. In theory, it should be easier to solve the detection problem. However, traditional encoder-decoders Structural Transformer do not work well in the densely predicted scenes, such as person recognition, target tracking and instance segmentation. The Twins-SVT Transformer, a hierarchical network structure similar to CNN, is perfectly adapted to this task.

The structure of Twins-SVT Transformer is similar to that of ResNet, as shown in Figure 1.

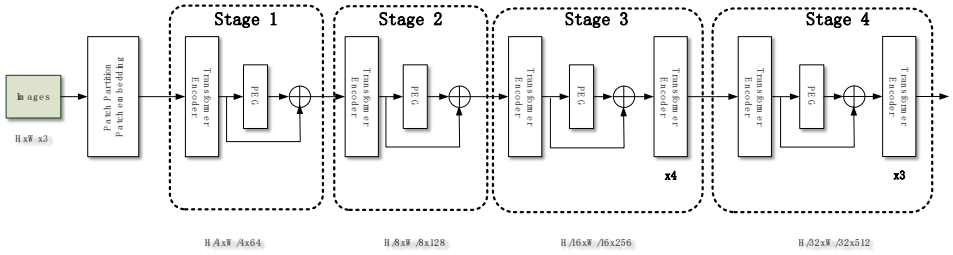

Figure 1. Twins-SVT model structure.

Patch Partition is to divide the original image into non-overlapping patch blocks, using the patch size of 4×4, so the dimension after Patch Partition is $(H/4 \times W/4 \times 48)$. The function of Patch embedding is to map the original feature to any dimension C. PEG is the position encoder in CPVT[17]. The structure of Transformer Encoder is shown in Figure 2.

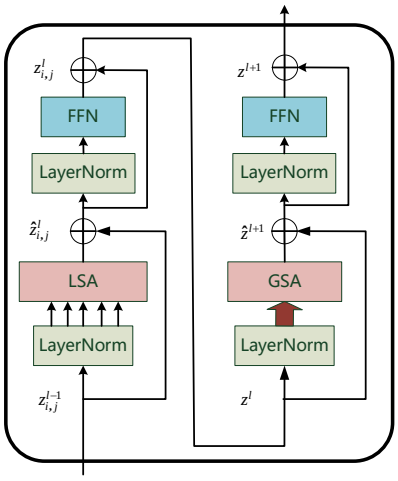

Figure 2. Twins Transformer Encoding structure.

The Transformer-Encoding in Figure 2 is the core point of the model. Based on a detailed analysis of the current global attention, this structure optimizes and improves the attention strategy. The new strategy integrates the local-global attention mechanism. The author compares it with the depth wise separable convolution [19] in the convolutional neural network, and named as Spatially Separable Self-Attention(SSSA). Different from the depth separable convolution, the SSSA proposed by Twins-SVT is to group the spatial dimensions of the features, calculate the self-attention of each group, and then integrate the grouped attention results from the global perspective.

Spatially Separable Self-Attention uses the local-global attention alternation (LSA-GSA) mechanism, which can greatly reduce the computational cost, and the complexity is reduced from the square order of the input to the linear order. By summarizing the attention of grouping calculation and using it as the key to calculate the global self-attention, the local attention can be transmitted to the global. The specific calculation method of SSSA is given by formula (3).

$$\begin{cases} \hat{z}_{ij}^{l\text{-}1} = LSA(LayerNorm(z_{ij}^{l-1})) + z_{ij}^{l-1}, \\ z_{ij}^{l} = FFN(LayerNorm(\hat{z}_{ij}^{l})) + \hat{z}_{ij}^{l}, \\ \hat{z}^{l+1} = GSA(LayerNorm(z^{l})) + z^{l}, \\ z^{l+1} = FFN(LayerNorm(\hat{z}^{l+1})) + \hat{z}^{l+1}, \\ i \in \{1,2,3,\ldots,m\}, j \in \{1,2,3,\ldots,n\} \end{cases} \tag{3}$$

Among them, z_{ij}^{l-1} represents the ij-th subwindow of block $l-1$, and z^{l} represents the output characteristics of the FFN (Feed Forword Network) module of block l.

Finally, after stage4 is completed, the dimension of the feature map is $(512,7,7)$, and the output layer of Twins-SVT Transformer uses a Global Average Pooling to obtain a feature vector of length 512 on this feature map, and then passes an LN and a full connection layer gets the final classification prediction vector.

3.3. Twins-SVT transformer

Because Twins-SVT Transformer is a multi-stage network framework and the output of each stage is also a set of feature maps, it can be easily migrated to almost all CV tasks. The author's experimental results also show that Twins-SVT Transformer has reached the state-of-the-art level in the field of detection and classification. A particularly important point in CNN is that as the network level deepens, the receptive field of nodes is also expanding. This feature is also satisfied in Twins-SVT Transformer. Therefore, this article will try Twins-SVT as the backbone of ReID. In order to assist the model to obtain more useful information from the target input and improve the accuracy of the model, this paper focuses on combining the feature information fusion between different stages of Twins-SVT.

The structure of the ReID model proposed in this paper is shown in Figure 3.

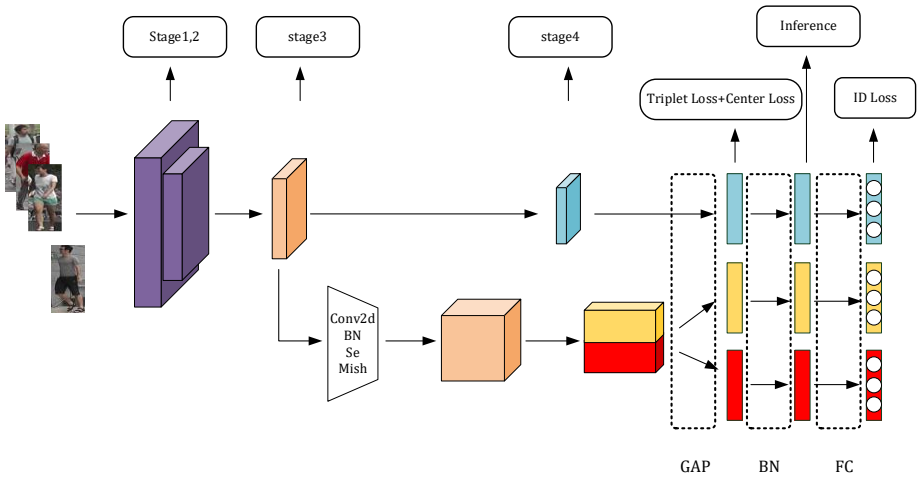

Figure 3. twinsReID model structure.

The first step is to modify the origin Twins-SVT Transformer model. Because the original Twins-SVT Transformer is used for image classification, the last adaptive pooling layer and head classification layer are removed. Then reshaping the feature whose original output dimension is 49 into a feature map f of 7×7. So the output of stage4 is the feature map of $b \times 512 \times 7 \times 7$. GAP is performed on the feature map of stage4, and the feature vector f_g is obtained after dimensionality reduction, and f_g' is obtained after the BN layer, the classification prediction vector p_g is obtained after the fully connected layer. This main branch is mainly used to extract the global characteristics of the character.

Considering that the feature map size of stage4 is only 7×7, some detailed features in the image may be lost. Therefore, we extracted the feature map of stage3 as auxiliary information for training, because the size of the feature map of stage3 is twice that of stage4, the feature map of stage3 will contain more detailed character information. We pass the output of stage3, a feature map of $b \times 256 \times 14 \times 14$ through a ConvBlock of a two dimensional convolution, BN, SE Attention[20] and Mish [21], an activation function. The function of two dimensional convolution is to reduce the size of the feature map, and at the same time expand the number of feature map channels to the same number of channels as stage4. The purpose is to achieve the consistency of the number of feature vector channels during joint training and facilitate joint Center loss training. The purpose of the BN layer is to normalize the feature map; SE attention is to increase the model's attention to key information; the Mish function is a smoother activation function than Relu, which increases the nonlinearity of the module, and can help the model training better improve the generalization ability of the model during gradient descent optimization. The role of ConvBlock is to get fine-grained features and increase the number of channels on the basis of retaining the feature information of stage3, so that the number of feature channels of stage3 is consistent with the number of channels of stage4.

After passing through a ConvBlock, a feature map of size $b \times 512 \times 12 \times 12$ was obtained. Two feature maps of size $b \times 512 \times 6 \times 12$ are obtained by horizontal segmentation of the feature map, and then two feature vectors f_{l1} and f_{l2} are obtained by GAP reduction. f_{l1}' and f_{l2}' are obtained through BN layer, and finally classification prediction vectors p_{l1} and p_{l2} are obtained through full connection layer respectively. In the training phase, the model is optimized by Triplet loss, Center loss and Cross entropy loss respectively. Among them, f_g, f_{l1} and f_{l2} are optimized by Triplet loss and center loss, and p_g, p_{l1} and p_{l2} are optimized by Cross entropy loss. According to BagofTricks [22], if ID loss and Triplet loss are used to optimize the same feature vector, the model may not be optimized to the optimal solution due to the inconsistency of the optimization goals of the two. Therefore, it is between the features optimized by ID loss and Triplet loss that added a BN layer. In the inference phase, the feature vectors after the BN layer are sequentially connected according to the channel dimensions as the feature representation of the input image, and the model inference and performance evaluation are performed.

4. Loss function

Person re-identification tasks are multi task learning in most cases. The mainstream is divided into two tasks. One is to construct ID Loss and learn losses corresponding to different ids by classifying losses. The purpose of ID loss is to model information in a specific field in order to distinguish different people in each mode. The other is a loss that is directly constructed through feature vectors based on Triple loss. It learns the similarity within the class and the distinction between classes so that the feature

vectors of different classes are more distinguishable, and the feature vectors of the same class are more similar.

4.1. Cross-entropy loss

In person re-identification tasks, the most commonly used ID loss is cross entroy loss, which describes the distance between two probability distributions. When the cross-entropy loss is smaller, the closer the two probability distributions are. For the input picture x, the prediction output of the model is denoted as p_i, which represents the predicted probability of the *i-th* class, and the true label of *x* is denoted as *y*, then the cross-entropy loss can be expressed as shown in formula (4).

$$L_{CE} = \sum_{i}^{N} -q_i \log(p_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases} \tag{4}$$

However, there are still a lot of negative samples in the process of ReID. The more classes, the greater number of negative samples. In order to build the loss of negative samples instead of ignoring negative samples, and to solve the possible over fitting problem of the model when training the classification loss and to improve the generalization ability of the model, the operation of label smooth [23] is introduced into the cross entropy. Unlike traditional cross entropy, the class label is not forced to be considered as 0 or 1, but is calculated with a certain probability. Compared with the traditional Cross entropy loss, its improvement lies in the smoother changes to q_i. The specific formula is shown in formula (5).

$$q_i = \begin{cases} 1 - \dfrac{N-1}{N}\varepsilon, & if \quad i = y \\ \dfrac{\varepsilon}{N}, & otherwise \end{cases} \tag{5}$$

Where N is the number of classes of the multi classification problem, and ε is a smaller hyperparameter.

4.2. Triplet loss

Due to the particularity of person re-identification tasks, the existence of metric learning is essential. Because the essence of ReID is to make the distance between same classes as small as possible and the distance between different classes as large as possible. Therefore, Triplet loss is used as the loss of metric learning. Triplet loss was originally proposed in the paper of FaceNet [24]. Similar images are similar in the embedding space, and can be judged whether they are the same one. The training goal is to make the samples with the same label as close as possible in the embedding space, and make the samples with different labels as far away as possible in the embedding space. Its calculation formula is shown in formula (6).

$$L_{Tri} = \left\lfloor d_p - d_n + \alpha \right\rfloor_+ \tag{6}$$

Among them, d_p and d_n respectively represent the distance of the feature representation pair

of the positive sample pair and the negative sample pair. Generally, the Euclidean distance is used to calculate it. α represents the margin distance of the loss, and [z]+ is equivalent to $max(z, 0)$.

In our experiment, α is set to 0.3. But the triple loss only considers the difference between d_p and d_n, and ignores their numerical value. For example, when $d_p = 0.3$ and $d_n = 0.5$, the triplet loss is 0.1. For another case, when $d_p = 1.3$ and $d_n = 1.5$, the triplet loss is also 0.1. It is difficult to ensure that d_p is smaller than d_n throughout the training process. In addition, Triplet loss does not take into account the compactness of the samples within the class.

4.3. Center loss

In order to make up for the lack of Triplet loss, we also added Center loss [25]. This loss constrains the class center of each class and minimizes the distance between each sample in the mini-batch and the corresponding class center, so that the distance within the class can be reduced and the compactness of the samples within the class can be achieved. The formula of the loss function is shown in formula (7).

$$L_C = \frac{1}{2 \times B} \sum_{j=1}^{B} \left\| x_i - c_{y_i} \right\|_2^2 \tag{7}$$

Where x_i is the extracted feature of the *i-th* picture in each batch, c_{yi} represents the class center of the $yi - th$ class described by the feature, and B is the size of the batchsize.

The features of each sample need to be obtained through a good network feature layer. As the calculation completed, the average value of the features of all samples is the class center c. Besides, a good network can only be obtained when the class center is added. The optimization process of Center loss is as follows:

Randomly generate c_{yi} and update c_{yi} in each batch. That is, the center is randomly initialized, and then the distance between the current data and the center is calculated in each batch, and then the distance in the gradient form is added to the center, similar to the gradient descent method, as shown in formula (8).

$$\frac{\partial L_c}{\partial x} = \frac{1}{B} \sum_{i=1}^{B} (c_{y_i} - x_i) \tag{8}$$

In order to prevent the jitter of the center during the optimization process, the gradient is multiplied by a γ factor, and the value of γ is between 0 and 1, our experiment uses γ as 0.5. As shown in formula (9).

$$\Delta c = \frac{\gamma}{B} \sum_{i=1}^{B} (c_{y_i} - x_i) \tag{9}$$

4.4. Final loss

The final loss of our model consists of the loss as in formula (10).

$$L_{total} = L_{CE} + L_{Tri} + \mu L_c \tag{10}$$

Among them, μ is the balance parameter of center loss. In our experiment, μ is set to 0.001.

5. More details

5.1. DataSet

The data sets used in this experiment are Market-1501 and dukeMTMC. The datasets are described as follows.

The Market-1501 dataset was collected on the campus of Tsinghua University, taken in the summer while constructed and made public in 2015. It includes 1501 ids and 32,668 instances captured by 6 cameras (including 5 high-definition cameras and 1 low-definition camera). Each instance is captured by at least 2 cameras, and there may be multiple images in one camera. The training set has 751 ids, containing 12936 images, and each person has an average of 17.2 training data. The test set has 750 ids, containing 19,732 images, and each person has an average of 26.3 test data. The person detection rectangles of the 3368 query images are drawn manually, while the person detection rectangles in the gallery are detected by the DPM detector. The fixed number of training sets and test sets provided by this data set can be used in single-shot mode or multi-shot test mode.

The dukeMTMC dataset is a large-scale labeled multi-target multi-camera person tracking data set. It provides a new large-scale high-definition video data set recorded by 8 simultaneous cameras, with more than 7000 single camera tracks and more than 2700 independent characters. dukeMTMC-reID is a person re-identification subset of the dukeMTMC data set, and provides a manually labeled bounding box. The data set statistics are shown in Table 1.

Table 1. Data set distribution statistics.

Datasets	Cameras	TrainIDs	TrainImgs	TestIDs	QueryImgs	GalleryImgs
Market-1501	6	751	12396	750	3368	19732
dukeMTMC-reID	8	702	16522	702	2228	17661

5.2. Evaluation index

The experimental evaluation indicators are mAP and Rank-k.

The full name of mAP is mean Average Precision, which means average precision. This indicator is the most commonly used evaluation indicator in multi-target detection and multi-label image classification. Because most of these tasks have more than one label, the common single-label image classification standard cannot be used. mAP is to sum up the AP (Average Precision) in the multi-classification task and then take the average.

Rank-k is the core index of ranking hit rate, which refers to the probability that the top k images in the search results (the highest confidence) have the correct results.

5.3. Experimental details

The experimental platform of this experiment is Linux 64-bit, and the graphics card used is RTX-8000. Load the Twins-SVT pretraining weights trained on ImageNet [26] onto the model, and use the Adam optimizer in the training phase to optimize, the base learning rate is $2.5 \times e^{-4}$, and the Warmup

learning rate [15] strategy is used. The batch size in the training phase is 64, that means each group has 4 ids and each id has 16 instances. In order to match the input of Twins-SVT Transformer, the input image is reshaped to a size of 224×224. After a total of 90 epochs of training, the model performance is evaluated.

6. Experimental results

This article mainly conducts experiments and comparisons on the impact of random erasure augment on the model, performance on the ReID data set, cross-domain experiment performance, multi-level feature ablation experiment and model parameters.

In order to improve the robustness and the adaptability of the model to the problem of target occlusion, during training, the input image is enhanced by random erasing augment(REA [27]). In REA, for mini-batch the probability of random erasing is p_e, while the probability of remaining unchanged is $1 - p_e$. When it gets random erasing, a region with size of (W_e, H_e) is randomly selected and the pixel value is erased with a random value. With other conditions unchanged, the impact of different p_e on model performance was tested on Market1501. The results are shown in Table 2.

Table 2. The impact of $\boldsymbol{p_e}$ on the performance of the model.

p_e	0	0.2	0.4	0.5	0.6	0.8
mAP	0.822	0.827	0.848	0.854	0.849	0.841

After using the best REA parameters, the results on Market1501 and dukeMTMC are shown in Table 3and

Table **4 错误!未找到引用源。**, and the cross-domain experiment results are shown in Table 5, where RK stands for reranking [28], and all results in the experiment are from single query.

Table 3. Market1501 Experimental results.

method	*mAP*	*R1*	*R5*	*R10*
OS Net(2019)[29]	0.849	0.948	-	-
HOReID(2020)[30]	0.849	0.942	-	-
Auto-ReID(2019)[31]	0.851	0.945	-	-
AlignedReID++(2019)[32]	0.791	0.918	-	-
AA Net-50(2019)[33]	0.825	0.939	-	0.986
CASN(PCB) (2019)[34]	0.828	0.944	-	-
Strong baseline(2020)[22]	0.859	0.945	-	-
MHN(2019)[35]	0.85	0.951	0.981	0.989
SPReID(2019)[36]	0.834	0.937	0.976	0.984
twinsReID	**0.854**	**0.937**	**0.979**	**0.986**
twinsReID(RK)	**0.936**	**0.949**	**0.978**	**0.984**

Table 4. dukeMTMC Experimental results.

method	*mAP*	*R1*
AlignedReID++(2019)[32]	0.697	0.821
OS Net(2019)[29]	0.735	0.886
HOReID(2020)[30]	0.756	0.869
Strong baseline(2020)[22]	0.764	0.864
CASN(PCB) (2019)[34]	0.737	0.877
AA Net-50(2019)[33]	0.726	0.864
MHN(2019)[35]	0.772	0.891
PSE+ECN(2019)[37]	0.757	0.845
twinsReID	**0.782**	**0.886**
twinsReID(RK)	**0.898**	**0.923**

Table 5. Cross-domain experiment results.

method	*M->D*		*D->M*	
	R1	mAP	R1	mAP
TJ-AIDL(2018)[38]	0.443	0.230	0.582	0.265
ATNet(2019)[39]	0.451	0.249	0.557	0.256
Strong baseline(2020)[22]	0.414	0.257	0.543	0.255
twinsReID without REA	**0.457**	**0.263**	**0.531**	**0.259**
twinsReID (RK) without REA	**0.482**	**0.384**	**0.566**	**0.379**

The traditional cross-domain ReID is essentially an unsupervised problem. Generally, a GAN [40] network generates a picture with the target domain style from the source domain picture, or finetunes the pseudo label generation of the target domain through a clustering method, so as to make the result achieve a better accuracy rate. After comparison, it is found that although the model in this paper does not make corresponding special treatment for cross-domain ReID, it also has better results than the experiments of some traditional cross-domain ReID papers. If cross-domain special processing is performed on the model, the results of the transformer in cross-domain ReID can be expected. Besides, it can be found that twinsReID (RK) has a much better performance than twinsReID. That's because rerank (RK) essentially reorders the k results identified by the algorithm so that the positive samples are further forward and the negative samples are further back. Therefore, compared with the rank-k score, the proportion of the first k samples with positive samples, the score of rank-k improves relatively.

At the same time, in order to verify the performance improvement of the multi-level features on the model, a comparative ablation experiment of the model was done on Market-1501. The comparison is between the model without Convblock and the model with ConvBlock, and the other experimental conditions are all controlled in the same way. The experimental results are shown in Table 6.

Finally, we compared the parameter of our transformer model with the traditional CNN model, mainly comparing the flops and params indicators of the model. The results are shown in Table 7.

Table 6. The impact of multi-level features on model performance.

Ablation experiment	*mAP*	*R1*	*R5*
No Convblock	0.770	0.904	0.964
With Convblock	**0.854**	**0.937**	**0.979**

Table 7. Model parameter size comparison.

model	*Input size*	*Flops(G)*	*Params(M)*
RRID(2019)[41]	$1 \times 3 \times 384 \times 128$	6.13	50.0
MSBA(2019)[42]	$1 \times 3 \times 384 \times 128$	6.71	26.1
MGN(2018)[43]	$1 \times 3 \times 384 \times 128$	11.92	70.4
Strong baseline(2020)[22]	$1 \times 3 \times 256 \times 128$	4.07	23.5
twinsReID	$1 \times 3 \times 224 \times 224$	**2.98**	**24.8**

The input sizes of different models in Table 7 are different because the backbones of different models have different requirements for the size of the input image. As you can see, in [22,41–43] the backbone of the model is CNN, and the backbone of twinsReID is Transformer, because Transformer does not use complicated convolution operations, so the flops and parameters of the model are much lower than those of the CNN models. Therefore, the transformer model architecture has certain potential and advantages in reducing the amount of model parameters and accelerating model inference.

7. Experimental results

7.1. Training metrics visualization

Visualize the training loss and training acc on Market1501, the result is shown in Figure 4, Figure 5 and Figure 6.

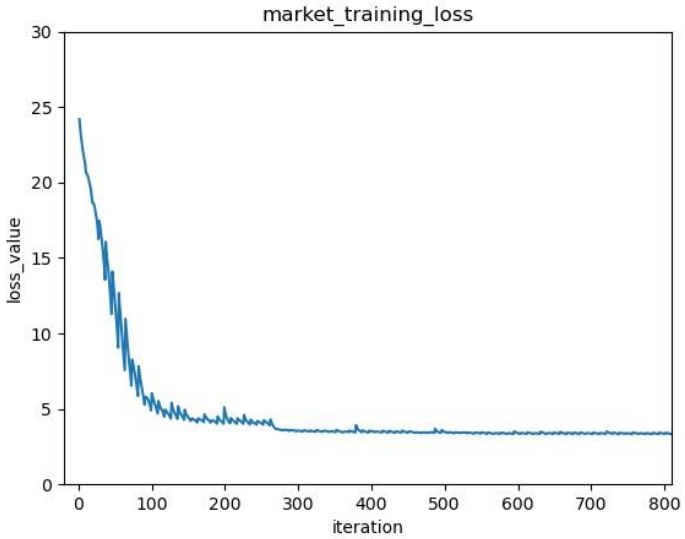

Figure 4. Loss visualization.

The Figure 4 is the change curve of total loss during the training process, the Figure 5 is the accuracy change of the training set during the training process, and the Figure 6 is the change curve of mAP and rank indicators on the test set.

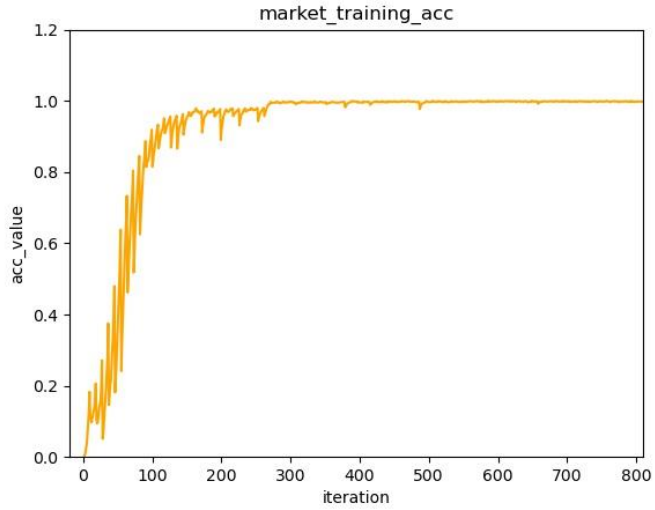

Figure 5. Acc visualization.

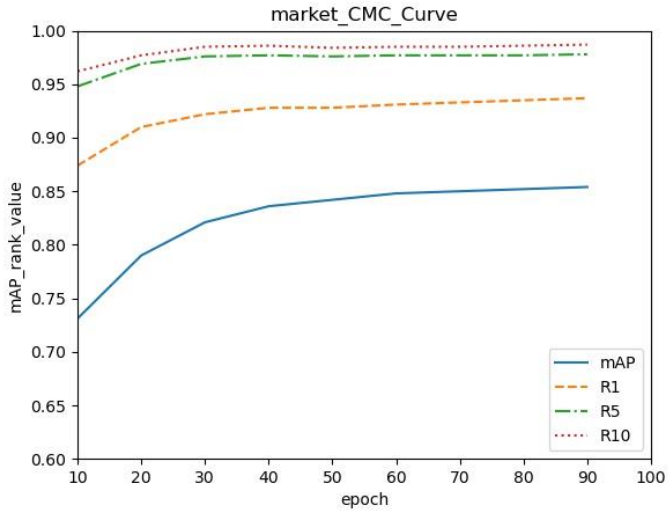

Figure 6. mAP and Rank visualization

7.2. Feature map visualization

In addition to the visualization of the training process, we also visualized the Twins-SVT Transformer to visualize the feature map of the ReID dataset with pretrained weights on ImageNet, as shown in Figure 7. At the same time, the visualization of the same image on ResNet50 is also visualized for comparison, as shown in Figure 8.

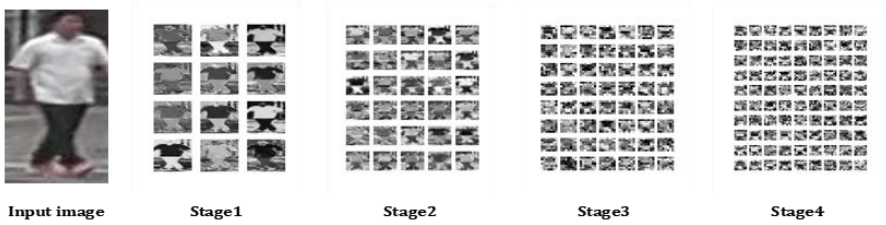

Figure 7. Visualization of feature maps of 4 stages of Twins-SVT.

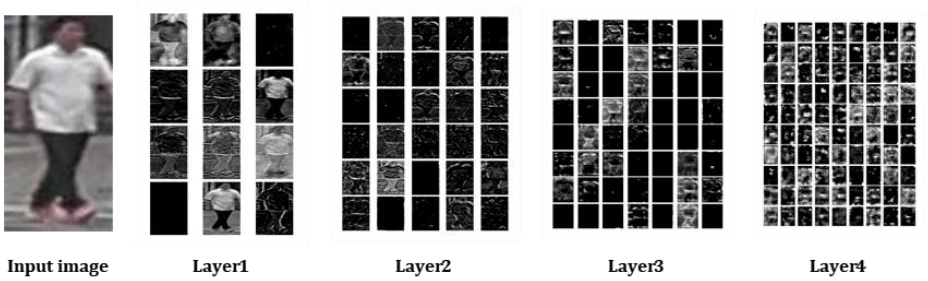

Figure 8. Visualization of feature maps of 4 stages of ResNet50.

Twins-SVT's stage1, stage2, stage3, and stage4 have 64, 128, 256, and 512 channels respectively, while in ResNet's layer1, layer2, layer3 and layer4, there are 256, 512, 1024, and 2048 channels respectively. In order to facilitate the display, only the characteristic diagrams of some of the channels are intercepted for display. It can be seen that Twins-SVT Transformer contains more and richer target feature information on a limited number of channels, so its image feature extraction ability is not inferior to the traditional CNN convolutional network.

7.3. Heat map visualization

To further illustrate the twinsReID response to person images, we made a heat map visualization of the model for some of the images of the Market1501, as shown in Figure 9. The highlighted part of the heat map shows that this part of the region occupies a relatively large proportion of the process of making classification and identification decisions for the model.

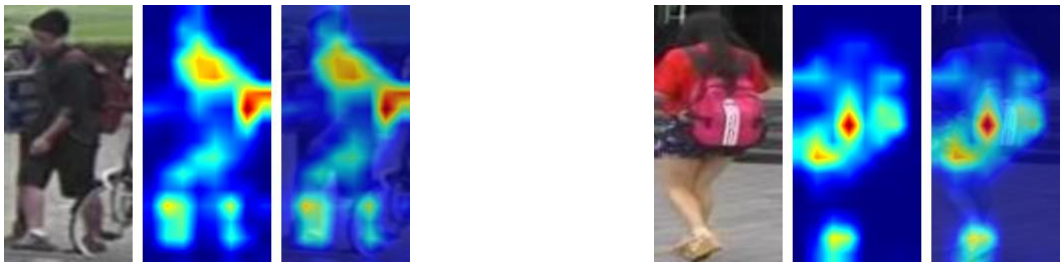

Figure 9. HeatMap of twinsReID.

7.4. Query results visualization

In single query mode, we show the query retrieval visualization on the market1501 data set, the result is shown in Figure 10.

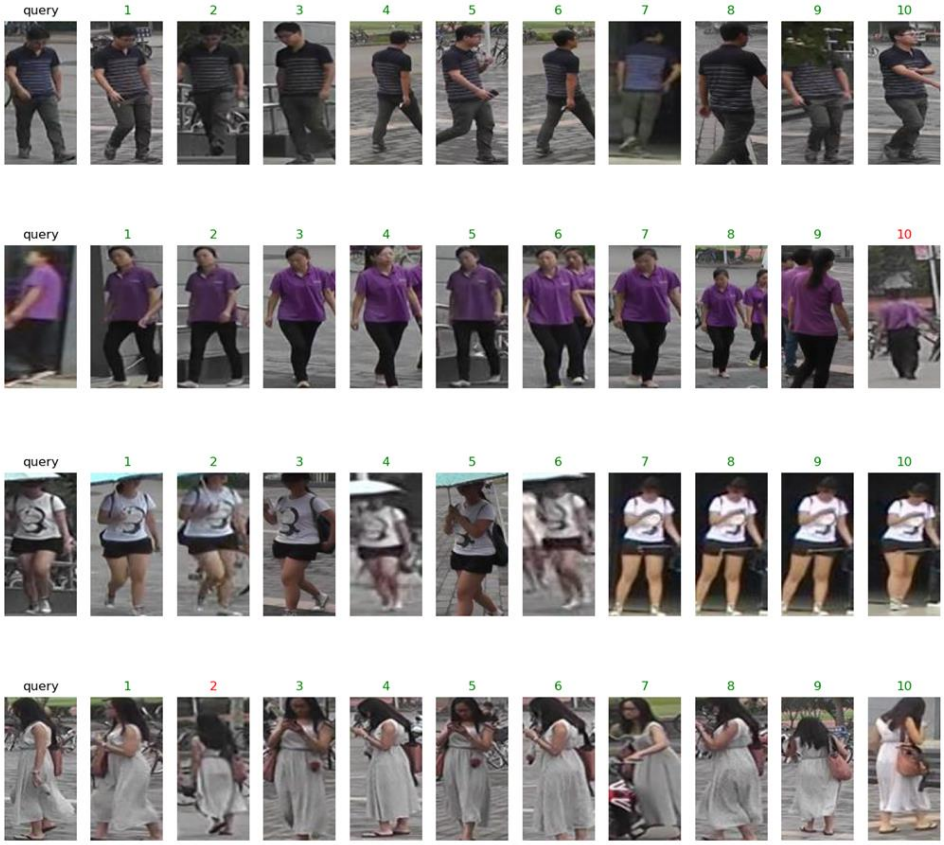

Figure 10. Visualization of query results.

The query of each subimage is the query image, and the 10 most similar target ids are queried. Green means the same id as the query, which means the query is correct, and red means it is not the same id as the query, which means the query is wrong.

8. Conclusion

Aiming at the problems of local receptive field and high computational complexity of CNN network in ReID task, this paper combined Transformer and ReID task, tried the feasibility of hierarchical Transformer model in ReID task, and proposed a novel ReID baseline model with multi-level features. The feature backbone of the model uses Twins-SVT instead of the traditional CNN architecture, and only uses Transformer for feature extraction. In the subsequent feature map processing phase, in addition to using the output features of the backbone, the feature map containing more detailed information from the previous stage of the backbone is also extracted. Through the convolutional layer with the attention mechanism, more fine-grained feature representations are obtained. Then feature stratification is carried out in order to get the response to the local detail features of the characters. Finally, joint training is carried out on the output features of the model.

The fusion training of features between different stages of Twins-SVT is the point explored in this article, and then the results of ablation experiments prove that the multi-level features are of great help to the improvement of the model's performance. Compared with the output feature map of ResNet50, which has 2048 channels, the number of channels of Twins-SVT is only 1/4 of that, although the Twins-SVT model in this paper achieves experimental results equivalent to ResNet50. Then compared the parameters of this model with the traditional CNN person re-identification models, which proves that the model has certain advantages in parameter lightening, and finally looks forward to the prospect of Transformer in cross-domain ReID.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (61971168, 62071161, 62271181), Zhejiang Provincial Natural Science Foundation of China (LZ22F010003).

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, *Neural Information Processing Systems (NIPS 2017)*, **29** (2017). https://doi.org/10.48550/arXiv.1701.04128
2. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, (2020), preprint. https://doi.org/10.48550/arXiv.2010.11929
3. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, (2018), preprint. https://doi.org/10.48550/arXiv.1810.04805
4. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, (2018), work in progress.
5. X. X. Chu, Z. Tian, Y. Q. Wang, B. Zhang, H. B. Ren, X. L. Wei, et al., Twins: Revisiting the design of spatial attention in vision transformers, *Neural Information Processing Systems (NIPS 2021)*, **34** (2021). https://doi.org/10.48550/arXiv.2104.13840
6. S. Cheng, I. C. Prentice, Y. Huang, Y. Jin, Y. K. Guo, R. Arcucci, Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting, *J. Comput. Phys.*, **464** (2022),111302. https://doi.org/10.1016/j.jcp.2022.111302
7. J. A. Weyn, D. R. Durran, R. Caruana, Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere, *J. Adv. Model. Earth Syst.*, **12** (2020). https://doi.org/10.1029/2020MS002109
8. M. Chen, A. Radford, J. Wu, H. W. Jun, P. Dhariwal, D. Luan, et al., Generative pretraining from pixels, *Proceed. Mach. Learn. Res.*, **199** (2020), 1691–1703.
9. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, *Proceed. Mach. Learn. Res.*, **139** (2021), 10347–10357. https://doi.org/10.48550/arXiv.2012.12877

10. K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, Y. Wang, Transformer in transformer, *Neural Information Processing Systems*, **34** (2021), 15908–15919. https://doi.org/10.48550/arXiv.2103.00112
11. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *European conference on computer vision (ECCV 2020)*, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
12. R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 7242–7252. https://doi.org/10.48550/arXiv.2105.05633
13. S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, TransReID: Transformer-based Object Re-Identification, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 14993–15002. https://doi.org/10.1109/ICCV48922.2021.01474
14. Z. Liu, Y. Lin, Y. Cao, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
15. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
16. W. Wang, E. Xie, X. Li, D. P. Fan, K. T. Song, D. Liang, et al., Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 548–558. https://doi.org/10.1109/ICCV48922.2021.00061
17. X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, et al., Conditional positional encodings for vision transformers, (2021), preprint. https://doi.org/10.48550/arXiv.2102.10882
18. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, *Neural Information Processing Systems (NIPS 2017)*, **30** (2017). https://doi.org/10.48550/arXiv.1706.03762
19. L. Sifre, S. Mallat, Rigid-Motion Scattering for Image Classification, (2014), preprint. https://doi.org/10.48550/arXiv.1403.1687
20. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 7032–7141. https://doi.org/10.1109/CVPR.2018.00745
21. D. Misra, Mish: A self regularized non-monotonic activation function, (2019), preprint. https://doi.org/10.48550/arXiv.1908.08681
22. H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2019), 1487–1495. https://doi.org/10.1109/CVPRW.2019.00190
23. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2818–2826. https://doi.org/10.1109/CVPR.2016.308
24. F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 815–823. https://doi.org/10.1109/CVPR.2015.7298682
25. Y. Wen, K. Zhang, Z. Li, Y. Qiao, A Discriminative Feature Learning Approach for Deep Face Recognition, in *European conference on computer vision (ECCV 2016)*, (2016).

https://doi.org/10.1007/978-3-319-46478-7_31
26. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., ImageNet large scale visual recognition challenge, *Int. J. Comput. Vision*, **115** (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y
27. Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in *Proceedings of the AAAI conference on artificial intelligence*, **34** (2020). https://doi.org/10.48550/arXiv.1708.04896
28. Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3652–3661. https://doi.org/10.1109/CVPR.2017.389
29. K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2019), 3701–3711. https://doi.org/10.1109/ICCV.2019.00380
30. P. Wang, Z. Zhao, F. Su, X. Zu, N.V. Boulgouris, HOReID: Deep high-order mapping enhances pose alignment for person re-identification, *IEEE Transact. Image Process.*, **30** (2021), 2908–2922. https://doi.org/10.1109/TIP.2021.3055952
31. R. Quan, X. Dong, Y. Wu, L. Zhu, Y. Yang, Auto-ReID: Searching for a part-aware ConvNet for person re-identification, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 3749–3758. https://doi.org/10.1109/ICCV.2019.00385
32. H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, C. Zhang, Alignedreid++: Dynamically matching local information for person re-identification, *Pattern Recogn. J. Pattern Recogn. Soc.*, **94** (2019), 53–61. https://doi.org/10.1016/j.patcog.2019.05.028
33. C.-P. Tay, S. Roy, K.-H. Yap, AANet: Attribute Attention Network for Person Re-Identifications, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7127–7136. https://doi.org/10.1109/CVPR.2019.00730
34. M. Zheng, S. Karanam, Z. Wu, R.J. Radke, Re-Identification With Consistent Attentive Siamese Networks, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 5728–5737. https://doi.org/10.1109/CVPR.2019.00588
35. B. Chen, W. Deng, J. Hu, Mixed High-Order Attention Network for Person Re-Identification, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 371–381. https://doi.org/10.48550/arXiv.1908.05819
36. M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 1062–1071. https://doi.org/10.1109/CVPR.2018.00117
37. M. S. Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 420–429. https://doi.org/10.1109/CVPR.2018.00051
38. J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 2275–2284. https://doi.org/10.1109/CVPR.2018.00242
39. J. Liu, Z.-J. Zha, D. Chen, R. Hong, M. Wang, Adaptive transfer network for cross-domain person re-identification, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7195–7204. https://doi.org/10.1109/CVPR.2019.00737

40. I. Goodfellow, J. Pouget-Abadie, M. Mirza, Conditional generative adversarial nets, in *Neural Information Processing Systems*, **27** (2014). https://doi.org/10.48550/arXiv.1411.1784
41. H. Park, B. Ha, Relation network for person re-identification, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2020). https://doi.org/10.48550/arXiv.1911.09318
42. H. Tan, H. Xiao, X. Zhang, B. Dai, S. M. Lai, Y. Liu, et al., MSBA: Multiple scales, branches and attention network with bag of tricks for person re-identification, *IEEE Access*, **8** (2020), 63632–63642. https://doi.org/10.1109/ACCESS.2020.2984915
43. G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in *Proceedings of the 26th ACM international conference on Multimedia*, (2018). https://doi.org/10.1145/3240508.3240552

AIMS Press