Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 5, No. 1, 2023

Kinship Verification Based on Global and Local Attention Mechanism

Decai Li[1, 2, a], Xingguo Jiang[1, 2, *]

1Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Zigong 643000, China
2School of Automation and Information Engineering, Sichuan University of Science and Engineering, Zigong 643000, China
[a]1340998024@qq.com, [*] **Corresponding author**: 750655632@qq.com

Abstract: Kinship verification is an important and challenging problem in computer vision. How to extract discriminative features is the key to improve the accuracy of kinship verification. At present, convolutional neural networks (CNNs) for feature extraction in the field of computer vision has achieved remarkable success, making it the most scholars used to study kinship verification related issues. However, few people use the self-attention mechanism with global capture capability to build a backbone feature classification network. Therefore, this paper proposes a backbone feature extraction network model based on a non-convolution, which expands the selection range of traditional classification networks for kinship verification related issues. Specifically, the paper proposes to use Vision Transformers as the basic backbone feature extraction network, combined with CNN with local attention mechanism, to provide a unique integrated solution in kinship verification. The proposed GLANet model is used for kinship verification and can verify 11 kinship pairs. The final experimental results show that in the FIW dataset, compared with the RFIW2020 challenge leading method, the proposed method has better verification effect in kinship, and the accuracy rate can reach 79.6 %.

Keywords: Kinship verification, ResNet50, Vision Transformers, Siamese neural network, Deep learning.

1. Introduction

Face images contain a large number of biological features. Existing psychological and sociological studies[1] have shown that face is an important clue to judge the similarity of kinship. Kinship verification is a new research topic in the field of computer vision, which is used to predict whether there is a kinship between a given face image.

There are significant differences between kinship verification and face recognition. Face recognition object is the same person, the same person's facial features in the short term little change. The face images used for kinship verification come from different people. These face images have great differences in posture, illumination, expression, age, gender, occlusion and other conditions. In addition, the complexity of genetic characteristics also caused a variety of facial appearance changes. At present, kinship verification algorithms are roughly divided into three categories: methods based on manual features [2-3], methods based on metric learning[4-5], and methods based on deep learning [6-8]. The feature-based method mainly uses artificial features and traditional classifiers to verify kinship. This method has low verification accuracy and manual extraction of features. Based on the metric learning method, a statistical learning method is used to learn an effective classifier or distance metric. The learned model can increase the distance between non-kinship pairs and reduce the distance between kinship pairs. Although the above two methods improve the accuracy of kinship verification, how to extract more distinguishable features from different kinship faces is the key to improve the accuracy. The method based on deep learning is a hot topic in current research. It mainly extracts the depth features of face images, analyzes the depth features, and obtains the verification results of kinship. The method based on deep learning further improves the accuracy, but most of the feature extraction backbone networks use CNN with local attention mechanism, and build the network around the VGGFace-Resnet50 architecture and some less commonly used CNN models. Few researchers use the Vision Transformers (ViT) model with global self-attention mechanism as the backbone feature extraction network.

Aiming at the above problems, this paper proposes a siamese neural network GLANet, which combines local attention CNN and a Vision Transformer with global attention mechanism. The Resnet50[9] and PVT[10] pre-trained models are used as the backbone of the twin neural network, and then the features extracted from the backbone model are 1×1 convolution and feature fusion. Finally, the generated features are input into the fully connected network for kinship verification. Experimental results show that this method can extract features of kinship more effectively and improve the accuracy of kinship verification. Compared with the leading method of the FG2020 Challenge, the twin neural network can achieve better results.

2. Feature Extraction Backbone Network

2.1. ResNet50 Feature Extraction Network

The residual neural network (ResNet) proposed by He Kaiming, Zhang Xiangyu and others [9] made the main contribution to the discovery of "degradation phenomenon", and a residual structure is invented for the degradation phenomenon, which greatly eliminates the difficulty of training neural networks with large depth. Subsequently, the network models such as ResNet-34, ResNet-50 and ResNet-101 are derived. The residual principle diagram is shown in Fig. 1.

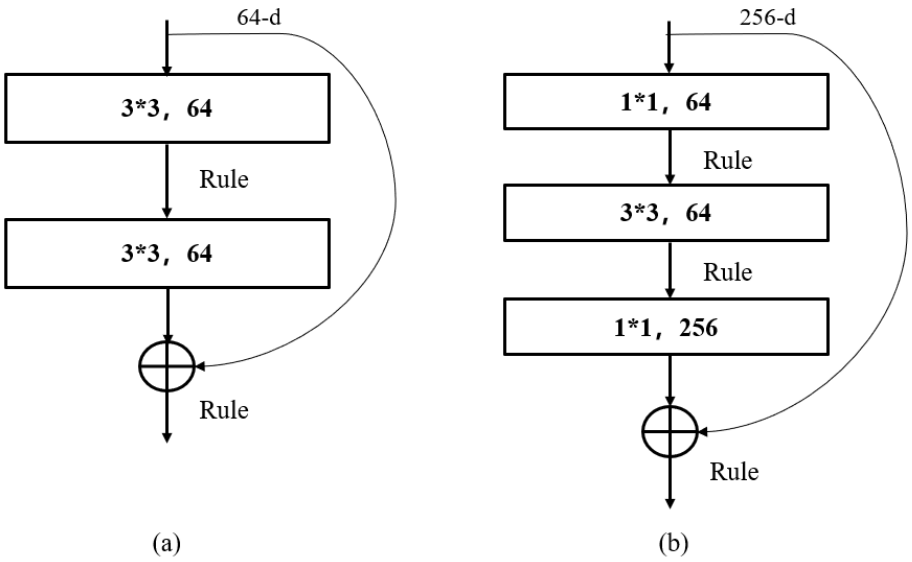

Figure 1. Residual structure

Among them, Fig.1(a) is a double residual module, the main branch is composed of two layers of 3×3 convolution layers, each of which is followed by Batch Normalization. Residual structure is a special kind of constant mapping, which does not introduce many parameters and calculation. The residual structure is composed of the direct output x of the right branch and the output $f(x)$ of the stack layer, and the finally output is expressed as:

$$y = f(x) + x \tag{1}$$

In the equation , x , y denote the input and output respectively, $f(x)$ denote the residual mapping.

The deeper ResNet50, ResNet101 and ResNet152 use three-layer residual structure. Fig.1(b) shows the three-layer residual module. The dimension's reduction & elevation are realized through two 1×1convolution layers, which can effectively solve the problem of excessive parameters of the two-layer residual module and time-consuming training. ResNet50 is selected as the backbone feature extraction network of twin neural network model. The detailed parameters of ResNet50 architecture are shown in Table 1.

Table 1. Detailed ResNet50 architecture parameters

layer name	Output size	ResNet50
Conv1	112×112	7×7，64，stride2
		3×3max pool，stride2
Conv2_x	56×56	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
Conv3_x	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$
Conv4_x	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$
Conv5_x	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, 1000-d fc

2.2. PVT Feature Extraction Network

PVT is an improved version of Vision Transformer proposed by Wang et al.[10], which has two main advantages: 1) It generates multi-scale feature maps between each block, combining the advantages of CNN. 2) It introduces Spatial-reduction technology to greatly reduce Transformer computation and memory to train each model. An overview of the PVT network architecture is shown in Figure 2.

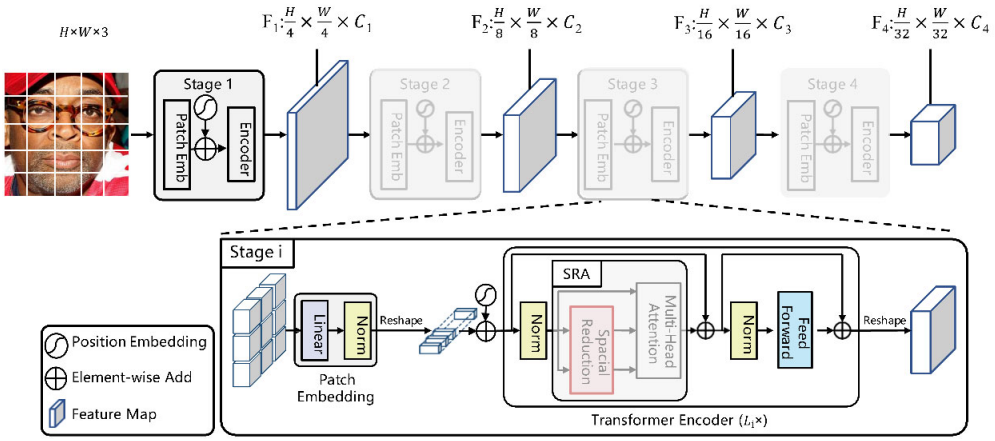

Figure 2. PVT backbone feature extraction network architecture

In PVT, the Transformer encoder in stage i has L_i encoder layers, each of which consists of an attention layer and a feedforward network layer. PVT is easier to train because it replaces the traditional multi-head attention layer[11] with a spatial-reduce attention (SRA) layer. This newly designed SRA layer performs similar to multi-head attention, which receives a query Q, a key K, and a value V as input and output characteristics. However, SRA reduces the spatial scale of K and V before the attention operation to reduce the computational overhead. The detailed description of SRA is as follows:

$$\mathrm{Reduce}(x) = \mathrm{Norm}\left(\mathrm{Reshape}\left(x, R_i\right) W^S\right) \quad (2)$$

Formula (2) describes how to spatially restore the input sequence. In the formula, x represents an input sequence, R represents the reduction ratio, W is a linear projection of the dimension reduction of the input sequence, and the input x is adjusted to $\frac{HW}{R^2} \times \left(R^2 C\right)$.

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{qk^T}{\sqrt{d_{\mathrm{head}}}}\right)v \quad (3)$$

$$\mathrm{head}_j = \mathrm{Attention}\left(QW^Q, \mathrm{Reduce}(K)W^k, \mathrm{Reduce}(V)W^v\right) \quad (4)$$

$$\mathrm{SRA}(Q,K,V) = \mathrm{Concat}\left(\mathrm{head}_0, \ldots, \mathrm{head}_{Ni}\right) \quad (5)$$

The rest of the calculation is the same as the original multihead attention. First, Attention is used to calculate the pairwise similarity between two elements of a sequence and their respective $Q$and K . W^Q , (K) W^k , W^v are the linear projection parameters of Q, K and V, respectively. N is the number of heads of attention layer in stage i. Then, the attention scores of each head are calculated and concatenated for the final SRA output. Therefore, SRA is a simple but effective attention layer that can process high-resolution feature maps while reducing computational and memory costs. The final output of PVT is a feature vector that can be input into the siamese network for downstream tasks of kinship verification.

3. GLANet Model and Loss Function

3.1. GLANet Model

ResNet50 network and PVT network are used to construct siamese branches to extract features. For ease of description, it is abbreviated as GLANet, and its architecture is shown in Figure 3. The two face images are extracted by two backbone twin networks respectively. The extracted features are used to reduce the feature dimension by feature fusion and 1× 1 convolution, and then combined and connected into a long vector. Then the long vector is input into the fully connected network Full Connection (FC) to measure the relative similarity between the two face images, and finally determine whether the two images have kinship. The FC layer consists of three fully connected layers, two relu activation functions and a sigmoid activation layer.

In order to better extract the relative face information and speed up the model training speed, the two feature extraction backbone networks are pre-trained on the large data set ImageNet[12] to narrow the appearance gap between the old face and the young face, and can obtain more semantic information of the face.

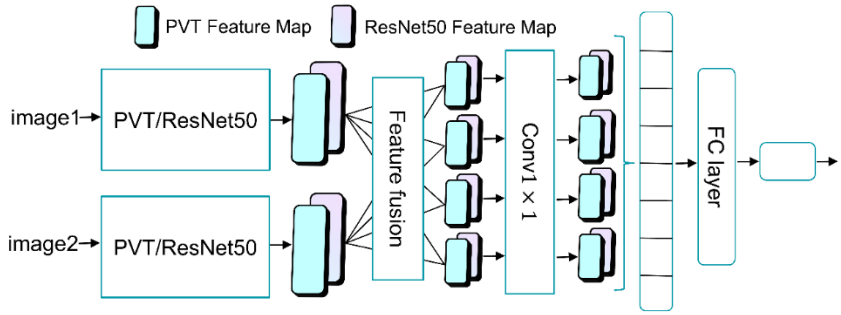

Figure 3. DSNN-TL Model Architecture

3.2. Loss Function

The Softmax function is mainly used to solve multi-classification problems. The results obtained by Softmax represent the probability that the input image is assigned to each category. Softmax loss is a loss function that combines Softmax and cross-entropy loss [13].

This paper proposes a joint supervised learning of Softmax loss and center loss [14], which reduces the intra-class distance while increasing the inter-class distance, so that the obtained features have stronger discrimination ability. As shown in (6):

$$L = -\frac{1}{m}\left[\sum_{i=1}^{m}\log\frac{e^{W_{y(i)}^{T}x(i)}}{\sum_{l-1}^{k}e^{W_{l}^{T}x(i)}}\right]+\frac{\lambda}{2}\sum_{i=1}^{m}\left\|x_i - c_{y_i}\right\|_2^2 \quad (6)$$

where m represents the number of samples, W is the weight of the network model, x represents the i-th image feature value, y_i represents the i-th class, and c_{y_i} represents the center of the classification feature value of the i-th image.

The center of the classification to which the picture belongs (the center of the eigenvalues of the classification); λ is used to balance two losses. Appropriate λ selection helps to enhance the feature discrimination ability of the network. When λ = 0, only Softmax loss supervised learning is used to train the network.

4. Experiment and Results

4.1. Experimental setup and parameters

The experimental environment is Windows 10, 64-bit operating system, 16GB memory, python programming language, TensorFlow deep learning framework, NVIDIA GeForce GTX 1650 and Intel (R) Core (TM) i7-10700F CPU@2.90 GHz (16 CPUs), ~2.9 GHz. Firstly, the experiment fine-tuned the two pre-training models, and all the parameters of ResNet50 and PVT feature extraction backbone network layer were frozen. The number of neurons in the second and third layers is 512 and 32 respectively. In order to prevent overfitting, dropout is introduced and set to 0.1. Then, the method of learning rate linearly decreasing with the training epoch is used to prevent excessive learning rate from oscillating back and forth when converging to the global optimum. After 10 epochs per iteration, the learning rate is attenuated by half when the maximum verification accuracy is not improved. The experimental parameter settings are shown in Table 2.

Table 2. Parameter settings

Parameter	Values
Epoch	100
BatchSize	16
Optimizer	Adam
Learning-Rate	0.0001

4.2. Experimental data set

The FIW [15] data set with the largest and most comprehensive facial image recognition by relatives is adopted, and the kinships included are divided into the following three types: peer relationship, brother-brother (B-B), sister-sister (S-S) and brother-sister (SI-BS) ; The first generation relationships: father and daughter (F-D), father and son (F-S), mother and daughter (M-D) and mother and son (M-S); The second generation: grandfather and granddaughter (GF-GD), grandfather and grandson (GF-GS), grandmother and granddaughter (GM-GD) and grandmother and grandson (GM-GS), a total of 11 pairs of kinship. Faces in FIW dataset contain changes in posture, facial expression, occlusion and lighting. Some images of some kinship faces are shown in Fig. 4.

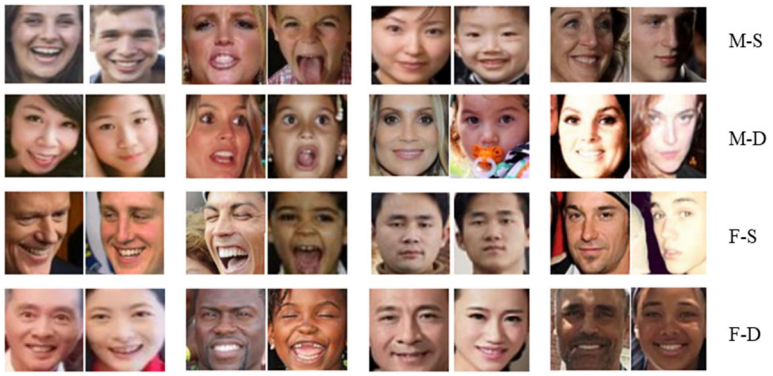

Figure 4. Face images of some relatives in FIW

4.3. Results and analysis

First, the joint loss function parameter λ is determined. The experimental results are shown in Figure 5. It can be seen from the figure that the superiority of the proposed Softmax loss and center loss joint supervised learning loss function, and its kinship verification accuracy is higher than that of only Softmax loss supervised learning (when λ = 0). At the same time, it can also be obtained from Fig.4 that a suitable λ value helps to improve the accuracy of kinship verification, and the best result of parameter λ is 0.003.

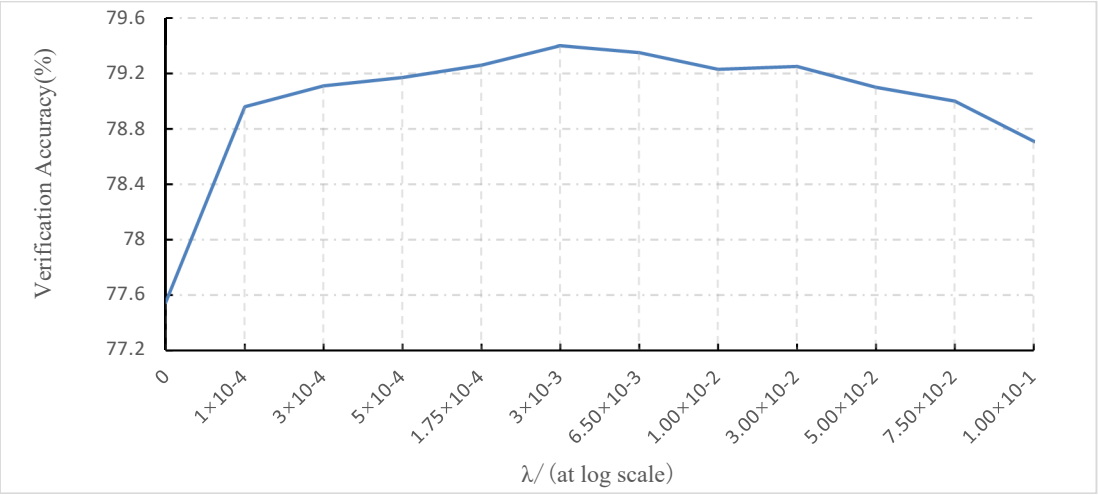

Figure 5. GLANet verification accuracy corresponding to different λ

Then, for the feature fusion method using CNN backbone network with good effect [6][7], the network accuracy of GLANet model under these different feature fusion methods is compared. The experimental results are shown in table 3. It can be seen from the table, the third group of $(x+y)^2$, $(x-y)^2$, $x \cdot y$, $_x{}^2 - y^2$ fusion methods obtained the highest accuracy of 79.6 %, and the subsequent GLANet model selected this group of fusion methods for experiments.

Table 3. Accuracy comparison of GLANet with different feature fusion methods

Feature Fusion	Acc (%)
$x+y$、$x-y$、$x \cdot y$、$1/2(x+y)$	78.5
$x+y$、$x-y$、$x \cdot y$、$_x{}^2 - y^2$	78.7
$(x+y)^2$、$(x-y)^2$、$x \cdot y$、$_x{}^2 - y^2$	79.6
$(x+y)^2$、$\sqrt{x}+\sqrt{y}$、$x \cdot y$、$_x{}^2 - y^2$	79.3

Table 4. Accuracy comparison with FG2020 Challenge leading method

Methods	F-D	F-S	M-D	M-S	GF-GD	GF-GS	GM-GD	GM-GS	B-B	S-S	SI-BS	Avg.
Baseline	0.61	0.66	0.69	0.62	0.66	0.71	0.73	0.68	0.57	0.64	0.5	0.64
Stefhoer	0.77	0.8	0.77	0.78	0.7	0.73	0.64	0.6	0.66	0.65	0.76	0.74
Ustc-nelslip	0.76	0.82	0.75	0.75	0.79	0.69	0.76	0.67	0.75	0.74	0.72	0.76
DeepBlueAI	0.74	0.81	0.75	0.74	0.72	0.73	0.67	0.68	0.77	0.77	0.75	0.76
Vuvko	0.75	0.81	0.78	0.74	0.78	0.69	0.76	0.6	0.8	0.8	0.77	0.78
GLANet	0.77	0.83	0.77	0.79	0.8	0.72	0.78	0.67	0.81	0.8	0.78	0.79

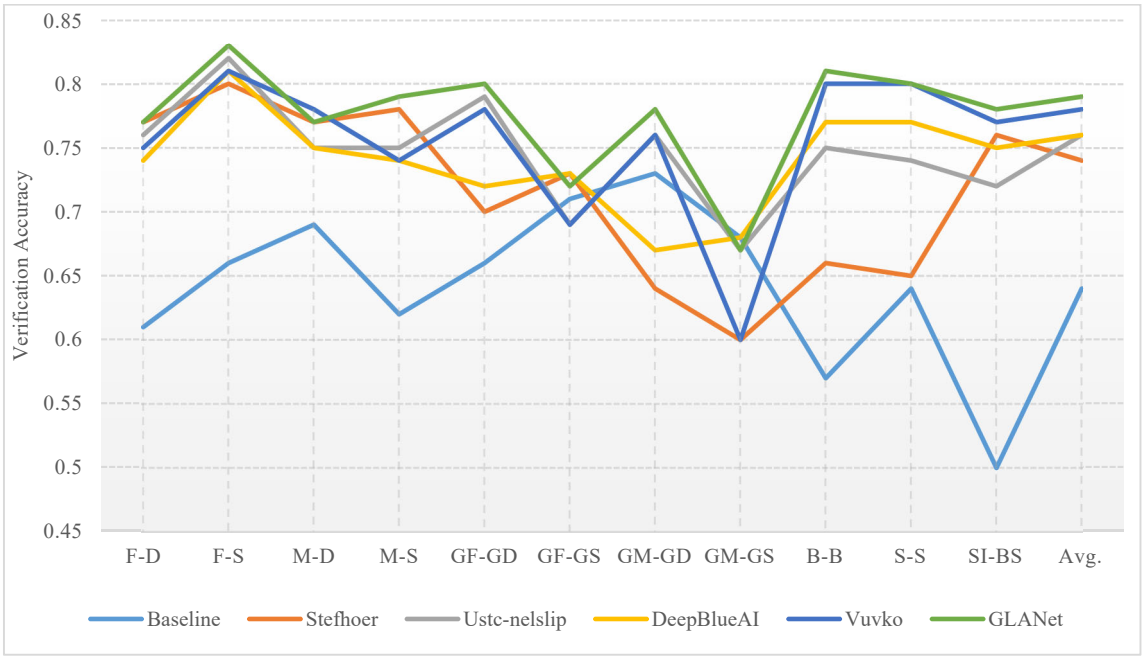

Figure 6. The validation accuracy line chart of 11 kinship pairs

In order to further verify the performance of the GLANet model, it is compared with the leading method of the FG2020 Challenge, and the results are shown in table 4. The experimental results show that the accuracy of GLANet is 15 % higher than that of the baseline model and 1.6 % higher than that of Vuvko, and the GLANet method obtains the best results. The line chart of the verification accuracy of 11 kinship pairs is shown in Fig.6.It can be seen that the proposed GLANet method can achieve the best accuracy compared with other methods in most relationship pairs. In GF-GS and GM-GS relation pairs, the accuracy of the methods listed is low, which is due to the small number of face images and age.

Finally, in order to further compare the feasibility and effectiveness of introducing ViT with global autonomous attention mechanism, that is, using PVT as the backbone feature extraction network, the common convolutional neural networks VGG16, SENet50 and Inception-ResnetV1 are used to replace the PVT backbone feature extraction network in the GLANet model, and all convolutional neural network models are pre-trained on ImageNet. Other settings are consistent with ResNet50 and PVT as GLANet feature extraction models for maximum accuracy. The results are shown in Table 5. Inception-ResnetV1 and ResNet50 have higher accuracy as backbone networks, reaching 77.8 %. However, the accuracy is 1.8 % lower than that of GLANet model, which proves the feasibility and effectiveness of the proposed method.

Table 5. Accuracy comparison using different backbone networks

Backbone	Acc
VGG16&ResNet50	74.2%
SENet50&ResNet50	76.7%
Inception-ResnetV1&ResNet50	77.8%

5. Conclusion

In this paper, a GLANet method for kinship verification is proposed. By introducing a ViT model with self-attention mechanism to make up for the shortcomings of the traditional convolutional neural network with local attention mechanism, a Siamese network is constructed to further fuse the facial features of relatives and further extract more discriminative features for kinship verification. The experimental results prove the feasibility and effectiveness of the proposed method. Compared with the leading method of FG2020 challenge, the accuracy of kinship verification is better, reaching 79.6 %.

Acknowledgment

This paper was supported by the Open Foundation of Artificial Intelligence Key Laboratory of Sichuan Province under Grant 2020RZJ03, and the Scientific Research Foundation of Sichuan University of Science and Engineering under Grant 2019RC12.

References

[1] DeBruine L M, Smith F G, Jones B C, et al. Kin recognition signals in adult faces[J]. Vision research, 2009, 49(1): 38-43.

[2] Fang R,Tang K D,Snavely N,et al.Towards computational models of kinship verification[C]//2010 IEEE International conference on image processing. IEEE, 2010: 1577-1580.

[3] YAN H,LU J,DENG W,et al.Discriminative multimetric learning for kinship verification[J].IEEE Transactions on Information Forensics and Security,2014,9(7): 1169-1178.

[4] ZHOU X Z,JIN K,XU M,et al. Learning Deep Compact Similarity Metric for Kinship Verification from Face Images [J]. Information Fusion,2019,48:84-94.

[5] LU J,ZHOU X,TAN Y P,et al.Neighborhood repulsed metric learning for kinship verification[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014,36(2):331-345.

[6] Dornaika F, Arganda-Carreras I, Serradilla O. Transfer learning and feature fusion for kinship verification[J]. Neural Computing and Applications, 2020, 32(11): 7139-7151.

[7] Robinson J P, Yin Y, Khan Z, et al. Recognizing families in the wild (RFIW): The 4th edition[C]//2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020: 857-862.

[8] Dahan E, Keller Y. A unified approach to kinship verification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(8): 2851-2857.

[9] Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 568-578.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[12] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[13] De Boer P T, Kroese D P, Mannor S, et al. A tutorial on the cross-entropy method[J]. Annals of operations research, 2005, 134(1): 19-67.

[14] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C]//European conference on computer vision. Springer, Cham, 2016: 499-515.

[15] Robinson J P, Shao M, Wu Y, et al. Visual kinship recognition of families in the wild[J]. IEEE Transactions on pattern analysis and machine intelligence, 2018, 40(11): 2624-2637.