Face Age Estimation With Multi-feature Fusion Model

Van Anh Nguyen[1,2][0000-0003-1711-6782]

[1] School of Computer Science Fudan University, Shanghai Key Laboratory of Intelligent Information Processing, No. 220, Handan Road, Shanghai, China 200433
[2] Viettel Group, Viettel Aerospace Institute, Thach That, Hanoi, Vietnam 13100
18210240272@fudan.edu.cn

Abstract. Age information is one of the most important features of human, so the task of extracting age features from face images has been received extensive attention, attracting many researchers. The appearance of a human face with the growth of age is affected by factors such as the difference of gender, race, environments… so, these tasks are of great significance and also brings challenges. In recent years, many researchers use deep learning techniques to solve the task of face age estimation. Specifically, use the VGG16 model to extract features, then use the classifier to estimate the age. However, disadvantage of this model which uses a lot of parameters, plus the depth of the network level so that slow model operation. Other researcher, use a rank-consistent ordinal regression method, using the ResNet34 structure to extract features, then combining the two-category extension method to achieve the age prediction task. This model has better results than the previous ordinal regression network on the UTKFace dataset but the MAE value is still large. To overcome the aforementioned shortcomings and improve accuracy, we have introduced a composite model that leverages various types of features, known as TransCNNFusion. The TransCNNFusion model combines the feature extraction abilities of the Attention mechanism with the local facial feature extraction of CNN. Experimental results demonstrate that the proposed model is as effective as or even superior to other Vision Transformer and CNN models, indicating its potential for practical applications.

Keywords: Estimate age; global feature; attention; local feature; CNN.

1 Introduction

The task of facial age estimation is an intriguing one, and it is approached differently by different researchers. Some consider it as a classification problem, where each age corresponds to a distinct category, while others view it as a regression task. In regression, age ranges are defined, and the goal is to predict which age range a facial image belongs to.

Facial images can contain various disguising factors such as facial hair, eyewear, makeup, etc., which can result in inaccurate feature extraction. Moreover, objective factors such as gender, race, environment, occlusions, and other challenges further complicate the extraction of facial age and gender features. Consequently, age estimation tasks pose significant challenges.

The availability of well-labeled datasets for age and gender is currently limited, and the collection and labeling process requires substantial human effort and time. Existing facial age datasets are not always ideal, as they may suffer from imbalanced distributions, small sample sizes, or insufficient age spans. These issues often lead to overfitting during experimentation, creating difficulties for developers.

In recent years, scientists have employed various methods and different approaches to tackle this problem, and they have achieved promising results. Notably, in 2018, Shen et al introduces a novel approach that combines deep learning and regression forests for accurate age estimation. By leveraging a deep neural network to extract facial features and training a regression forest model, the proposed method achieves state-of-the-art performance in age estimation tasks, making significant contributions to the field of machine vision [1].

Then, 2019, Cao et al introduces a novel approach that incorporates rank consistent ordinal regression into neural networks for improved age estimation accuracy [2]. The research demonstrates that this integration effectively addresses bias and inconsistency problems in age data, resulting in significantly enhanced precision in age estimation tasks. Additionally, Zhang et al. introduced the C3AE method [3], presents a novel approach that explores the boundaries of compact model design for age estimation. The proposed model offers an innovative and advanced compact architecture that has the potential to tackle age estimation challenges, particularly in scenarios with limited computational resources.

In 2020, Huynh et al. aimed to develop a method that combines age estimation and gender classification on Asians faces, using the Wide ResNet method [4]. In 2021, Deng et al. introduces a method that focuses on multi-feature learning and fusion to enhance the accuracy of age estimation from facial images. The proposed approach incorporates a combination of geometric, color, and texture features, aiming to improve the overall performance of age estimation models [5]. In 2022, Bonet Cervera et al. introduced a deep neural network (DNN) capable of performing multiple tasks, including predicting facial age [6].

The Transformer is a classical model widely applied in the field of natural language processing [7, 8]. Recently, researchers have explored its application in computer vision tasks, and surprisingly, it has yielded remarkable effectiveness. The success of the Vision Transformer (ViTs) has paved the way for a new research direction [9], making Transformer one of the preferred methods for addressing classification and image processing problems. Notably, the Transformer Model is not confined to NLP but extends its applicability to Computer Vision. Within image processing, the Transformer model can execute tasks like recognition, classification, image generation modeling, and even facial age estimation. In 2023, M. Kuprashevich et al introduced the MiVOLO (Multi Input VOLO) model for age and gender estimation [10].

While Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) models utilize convolutional mechanisms to extract local features from images, the Transformer model employs attention mechanisms to extract global features. Each approach can extract distinct types of features, but is there a method to extract both types of features simultaneously? To answer this question, we propose the TransCNNFusion model, which combines Convolution and Attention mechanisms to extract facial features, thereby addressing the task of facial age prediction from portrait images.

The main contributions of our research are as follows:

- Present a novel 2-branch Transformer model that enables the extraction of multi-scale features for facial age estimation. The model incorporates the FusionAttention mechanism, which combines Softmax Attention and Linear Attention to effectively extract features. Additionally, we design a branch of Convolutional Neural Networks that leverages convolutional mechanisms to extract local facial features. By combining these different types of facial age features, we provide an estimated age prediction.
- Apply the sharpness-aware minimizer (SAM) optimization method to the model when conducting training and optimization, thereby improving the accuracy of the model.
- Our model has comparable or better performance than CNN models and MiVOLO models.

2 Methods

2.1 Main structure of TransCNNFusion

The main structure of TransCNNFusion model is described in **Fig. 1**. The model takes a face image as input and processes it through three main branches: Trans2, CNNs, and Trans4. In these branches, Trans2 and Trans4 utilize the Transformer Encoder block, which employs the FusionAttention mechanism to extract features and optimizes the system using the internal SAM method. The CNNs branch consists of multiple layers, with each layer composed of several identity blocks and convolution blocks. The convolutional mechanisms are used to extract features.

The first branch – Trans2, will divide the input image into 4 small patches (2x2), add relative position parameters for the small patches, and then perform mapping to obtain input of the Transformer Encoder block. Positional encoding equips the model with the capability to consider all spatial information and location information of all patches.

The second branch – CNNs, this branch consists of multiple layers, with each layer comprising an identity block and multiple convolution blocks. It combines the use of convolution blocks and identity blocks to extract local facial features.

The third branch – Trans4, will split the input image into 16 smaller patches (4x4), then add relative position parameters for the small patches, and finally perform mapping to obtain input of the Transformer Encoder block.

The Transformer Encoder block consists of several sub-blocks, including FusionAttention block, and MLP block.

FusionAttention: We utilize a combination of Softmax attention (self-attention mechanism) and Linear Attention to analyze the interaction of patches in space. FusionAttention enables the model to identify significant relationships between patches and extract crucial information.

Residual Connections and Layer Normalization: To avoid the vanishing gradient problem and help the model converge faster, we use residual connections and layer normalization after each layer of the Transformer Encoder. Residual connections allow information to pass through layers without too much variation, while layer normalization helps stabilize the learning process.

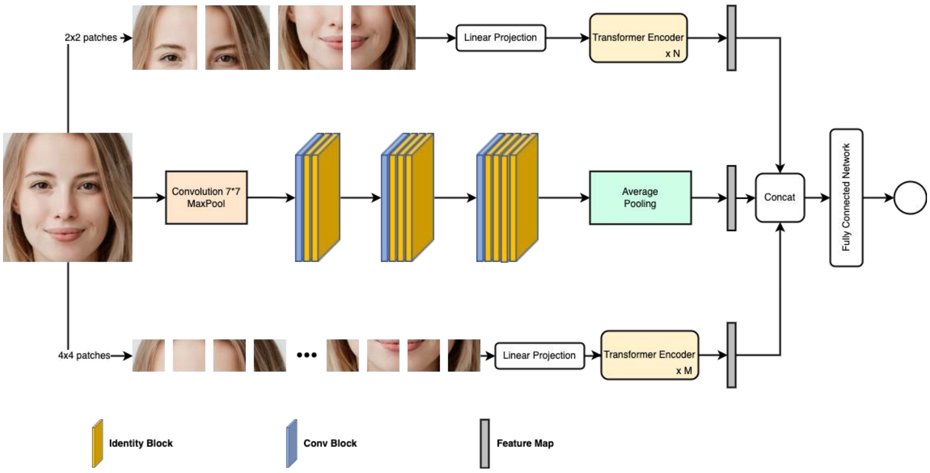

Fig. 1. Overall structure of TransCNNFusion

From the features extracted from the three branches, we proceed to merge them and pass them through the two fully connected layers (FCN) and the Softmax function to obtain the final output, thereby providing the estimated age result.

2.2 Image Preprocessing

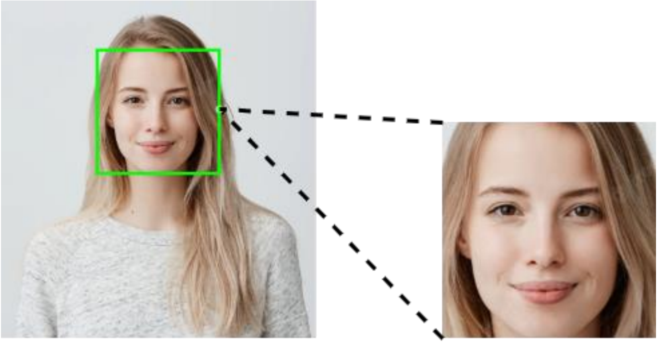

Fig. 2. Image preprocessing details: Extracting the area containing the face through the location determined from the image

The primary aim in developing this model is to create a comprehensive system that can receive an input image containing a face and accurately estimate the age of the person depicted. To accomplish this goal, we employ the [11]library, a widely-used open-source resource, for face detection and localization . Following successful detection and localization, we extract the identified faces and save each face as an individual file for further processing.

This crucial preprocessing step significantly influences the expected results of the model, and the process is detailed in **Fig. 2**.

2.3 Transformer Encoder Block

The Transformer Encoder block consists of several sub-blocks, including FusionAttention block, Norm layer, and MLP block.

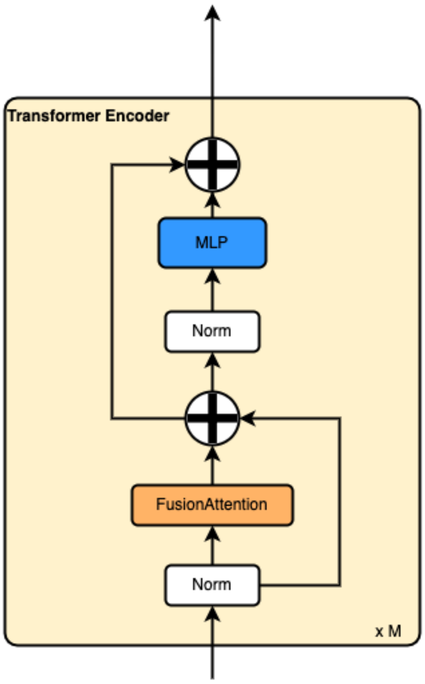

Fig. 3. The internal structure of Transformer Encoder block

The Transformer block receives a 1D vector as input and utilizes the FusionAttention mechanism to extract global features of face with a length of 1536. In the Trans2 branch, the Transformer Encoder block is iterated N times, while in the Trans4 branch, the Transformer Encoder block is iterated M times. The specific structure of the Transformer Encoder is illustrated in **Fig. 3**.

When calculating the Attention function, a set of queries is simultaneously packed into a Q matrix, keys and values are also packed together into a K matrix and a V matrix.

Softmax Attention calculates the similarity between all query – key pairs, it is calculated according to the following formula:

$$\mathrm{SoftmaxAttention} = \mathrm{Softmax}(\mathrm{QK^T})\mathrm{V} \tag{1}$$

Linear Attention uses the mapping function ϕ(.) to Q and K respectively to change the calculation order, it is calculated according to the following formula:

$$\mathrm{LinearAttention} = \ \phi(\mathrm{Q})(\phi(\mathrm{K^T})\mathrm{V}) \tag{2}$$

FusionAttention is combined by Softmax Attention and Linear Attention, its calculation formula is as follows:

$$\mathrm{FusionAttention} = \mathrm{Softmax}(\mathrm{SoftmaxAttention} + \mathrm{LinearAttention})$$

$$\mathrm{FusionAttention} = \ \mathrm{Softmax}(\mathrm{Softmax}(\mathrm{QK^T})\mathrm{V} + \phi(\mathrm{Q})(\phi(\mathrm{K^T})\mathrm{V})) \tag{3}$$

The detailed structure of FusionAttention block is shown in **Fig. 4**.

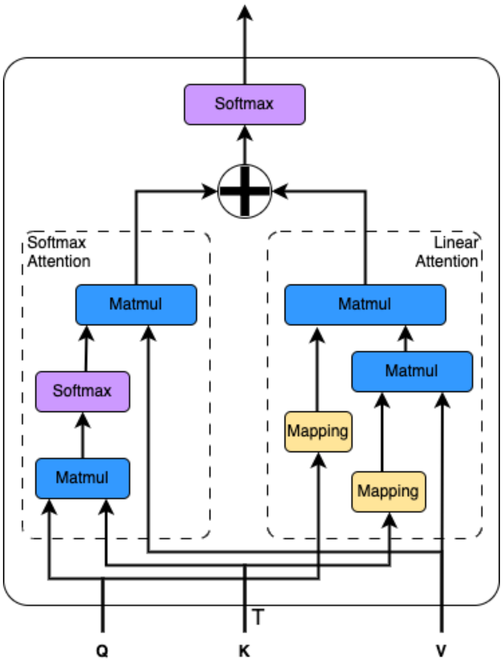

Fig. 4. The internal structure of FusionAttention block

2.4 CNNs Encoder

Within the CNNs branch, the input consists of a face image with dimensions 224 * 224, converted to a tensor data type. We employ Conv block and Identity block for face feature extraction, resulting in a vector of length 2048. These blocks are designed based on the network-in-network (NiN) structure. In comparison to conventional CNN structures like AlexNet and VGG, the NiN structure utilizes significantly fewer parameters for feature extraction, leading to faster training of the algorithm. The Conv block and Identity block primarily employ convolution operations with kernel sizes of 1x1 or 3x3, coupled with the ReLU activation function.

The specific structure of these two block types is depicted in **Fig. 5**.

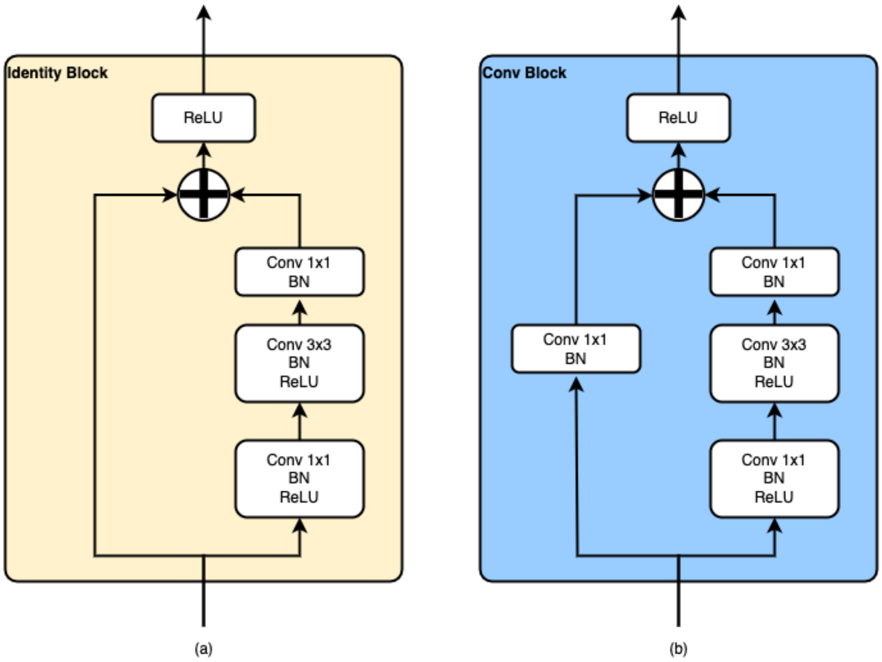

Fig. 5. The internal structure of Identity block (a) and Conv block (b)

2.5 SAM Optimization

The sharpness-aware minimizer (SAM) is a method introduced by Foret et al, SAM aims to address the generalization gap issue in deep learning models [12].

SAM introduces a regularization term into the loss function during the optimization process, which modifies the gradient updates to drive the model to converge towards sharper minima. This technique aims to enhance the model's generalization performance by optimizing for sharpness. By doing so, SAM improves the model's ability to handle unseen data and enhances its overall generalization capabilities.

Experimental results have demonstrated the effectiveness of SAM in reducing the generalization gap and improving the performance of deep learning models. Notably, SAM has been shown to enhance the accuracy of the ViTs model without the need for additional data [13].

3 Experiment

3.1 Dataset

Dataset is one of the important parts of deep learning, an efficient model that cannot separate factors such as data set richness, balance and data allocation of the dataset. We use 3 datasets to conduct experiments, including: UTKFace dataset [14], Adience dataset [15], and FGNET dataset [16].

The Adience dataset is also extensive, consisting of over 26,000 portrait images capturing over 2,200 individuals. It is annotated with gender and age labels and has been divided into various subsets.

The FGNET dataset contains portrait images of individuals ranging from 0 to 96 years old. This dataset was collected from 82 individuals, with multiple photographs taken of each person throughout their aging process. In total, the FGNET dataset includes more than 1,000 portrait images.

The UTKFace dataset is a substantial dataset comprising of more than 20,000 images annotated with age, race, and gender information. The age range spans from 1 to 116 years, and the dataset maintains a balanced gender distribution, with women accounting for 48% of the samples.

3.2 Experiment Setup

Despite utilizing the SAM optimization method, we also employed techniques to augment the facial data.

When conducting experiments with UTKFace dataset and Adience dataset, we divided them into three subsets: the training set, which accounts for 65%; the validation set, which accounts for 20%; and the testing set, which accounts for 15%. We utilized FGNET, despite its smaller size, exclusively as a testing set.

The input image after undergoing preprocessing will be resized to 224x224 before performing model training. We use the CrossEntropyLoss loss function to evaluate the quality of the model and the SAM optimizer to update and optimize the parameter values of the model.

3.3 Performance Evaluation

To evaluate the effectiveness of the methods, we use the evaluation criteria MAE (Mean Absolute Error) and CS (Consistency Score). In which, the smaller the MAE value, the more efficient the model is; The larger the CS@k value, the higher the accuracy of the model gets.

In the TransCNNFusion model, the Trans2 branch performs operations to process four small patches into a 1-dimensional vector, serving as input to the Encoder block. Similarly, the Trans4branch processes sixteen small patches into a 1-dimensional vector for input to the Encoder block. The Transformer Encoder block employs the FusionAttention mechanism to extract features and utilizes the internal SAM method to optimize the system.

To demonstrate the superior performance of the TransCNNFusion model when combining the aforementioned factors, we present modified models derived from the TransCNNFusion model. The information of the model architectures of TransCNNFusion and ViTsAge, as shown in **Table 1**, describes whether the models use Trans2, Trans4 and CNNs. Additionally, the Transformer encoder employs which attention mechanism and utilizes which optimizer method. ViTsAge is an improved version of the ViTs model that incorporates certain modifications to cater to the task of facial age estimation; in the Attention column, FA stands for FusionAttention.

Table 1. Model Architectures Of TransCNNFusion And ViTsAge

Model	Trans2	Trans4	CNNs	Attention	Optimizer
TransCNNFusion-2	Yes	-	Yes	FA	SAM
TransCNNFusion-4	-	Yes	Yes	FA	SAM
TransCNNFusion-S	Yes	Yes	Yes	Softmax	SAM
TransCNNFusion-L	Yes	Yes	Yes	Linear	SAM
ViTsAge	-	Yes	-	Softmax	Adam
TransCNNFusion-A	Yes	Yes	Yes	FA	Adam
TransCNNFusion	Yes	Yes	Yes	FA	SAM

Based on the experimental results presented in **Table 2**, it is evident that using Trans4 is more effective than using Trans2 alone. Additionally, utilizing either Trans2 or Trans4 independently is less effective compared to employing both simultaneously in the mentioned branches. When utilizing the same set of parameters, the TransCNNFusion model, which incorporates the SAM optimization method, achieves higher accuracy compared to the Adam method used in ViTsAge. The Softmax Attention mechanism significantly enhances feature extraction in the Transformer Encoder block, particularly when combined with the Linear Attention mechanism. This combination proves to be more effective in facilitating feature extraction by the encoder block.

Table 2. Comparison Of Accuracy Of TransCNNFusion Model And TransCNNFusion Modified Models

Model	UTKFace MAE	UTKFace CS@5	Adience MAE	Adience CS@5	FGNET MAE	FGNET CS@5
TransCNNFusion-2	4.91	67.61	5.35	63.21	3.64	73.37
TransCNNFusion-4	4.37	68.71	4.62	64.29	3.43	73.92
TransCNNFusion-S	4.31	69.34	4.51	65.57	3.24	74.39
TransCNNFusion-L	4.88	67.65	5.32	63.53	3.73	73.63
ViTsAge	4.59	68.16	4.99	64.13	3.54	74.03
TransCNNFusion-A	4.09	70.63	4.36	66.04	2.88	76.15
TransCNNFusion	3.79	75.31	4.08	70.65	2.49	81.61

We conducted a comparative analysis of the TransCNNFusion model's accuracy with other widely recognized CNN models, namely InceptionV3 [17], ResNet50 [18], Cao et al. [2], IncepRes, as well as ViTs models such as ViTsAge and MiVOLO. This evaluation was performed on three distinct datasets: UTKFace, Adience, and FGNET. IncepRes is a model we developed by merging the InceptionV3 and ResNet50 models.

Table 3. Analysis Results Using Different Models On UTKFace Dataset

Model	MAE	CS@0	CS@5	CS@10
InceptionV3	5.42	14.33	63.66	85.37
ResNet50	5.49	15.71	64.08	86.13
Cao et al [2]	5.39	-	-	-
IncepRes	5.34	16.03	65.07	87.32
ViTsAge	4.59	16.96	68.16	88.67
MiVOLO	3.70	-	74.16	-
TransCNNFusion	3.79	21.10	75.31	91.45

The outcomes of the model evaluations on the UTKFace dataset are presented in **Table 3**. It is apparent that our model exhibits superior performance, as indicated by lower MAE and higher CS, in comparison to the CNN models. The accuracy of our model is comparable to that of MiVOLO and significantly surpasses ViTsAge.

Table 4. Analysis Results Using Different Models On Adience Dataset

Model	MAE	CS@0	CS@5	CS@10
InceptionV3	5.72	14.12	60.61	83.66
ResNet50	5.91	14.71	60.38	83.31
IncepRes	5.58	15.11	61.02	84.21
ViTsAge	4.99	15.68	64.13	86.23
MiVOLO	-	-	68.69	-
TransCNNFusion	4.08	18.51	70.65	89.35

The accuracy results of the model evaluations on the Adience dataset are presented in **Table 4**. In this dataset, the TransCNNFusion model exhibits higher accuracy compared to MiVOLO. Furthermore, the TransCNNFusion model outperforms both the CNN models and the ViTsAge model in terms of MAE and CS.

Table 5. Analysis Results Using Different Models On FGNET Dataset

Model	MAE	CS@0	CS@5	CS@10
InceptionV3	4.92	17.65	71.59	85.86
ResNet50	4.41	17.92	71.86	86.01
IncepRes	4.35	18.25	72.39	86.42
ViTsAge	3.54	21.89	75.62	90.29
TransCNNFusion	2.49	26.22	81.61	95.35

FGNET is a small-sized dataset consisting primarily of portrait images of individuals at different ages. When evaluating the accuracy of the models on this dataset, all models demonstrate fairly good results. The specific outcomes can be observed in **Table 5**.

We utilized the gender labels available in the UTKFace dataset to investigate the impact of gender on facial age prediction outcomes.

Table 6. Analysis of the influence of gender on different models

Method	UTKFace (Only Male)	UTKFace (Only Female)	UTKFace (Total)
InceptionV3	5.09	5.78	5.42
ResNet50	5.03	5.98	5.49
IncepRes	4.98	5.73	5.34
AgeTrans	3.86	5.38	4.59
TransCNNFusion	3.18	4.45	3.79

The results in **Table 6** indicate that the models tend to predict age more accurately for images with male gender.

4 Conclusions

We propose the TransCNNFusion model, which is designed with three different input branches utilizing the attention mechanism and convolutional mechanisms. Specifically, two branches employ a combination of Softmax Attention and Linear Attention, while one branch utilizes convolutional mechanisms for facial feature extraction. The model is able to extract both global and local facial features, thereby providing a viable solution to the challenge of predicting facial age. At the same time, the proposed model uses the SAM method to optimize the model, overcome the disadvantage of needing a large dataset when training the ViTs model, and improve the model's accuracy.

Based on the experimental results, the TransCNNFusion model demonstrates high accuracy on both the UTKFace dataset, the Adience dataset, and FGNET dataset. Moving forward, our focus will be on further research and enhancements to the model, aiming to achieve even higher accuracy.

References

1. Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., Yuille, A.L.: Deep regression forests for age estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2304-2313. (2018)

2. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters 140, 325-331 (2020)

3. Zhang, C., Liu, S., Xu, X., Zhu, C.: C3AE: Exploring the limits of compact model for age estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12587-12596. (2019)

4. Huynh, H.T., Nguyen, H.: Joint age estimation and gender classification of Asian faces using wide ResNet. SN computer science 1, 284 (2020)

5. Deng, Y., Teng, S., Fei, L., Zhang, W., Rida, I.: A multifeature learning and fusion network for facial age estimation. Sensors 21, 4597 (2021)

6. Bonet Cervera, E.: Age & Gender Recognition in The Wild. Universitat Politècnica de Catalunya (2022)

7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30, (2017)

8. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp. 38-45. (2020)

9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

10. Kuprashevich, M., Tolstykh, I.: MiVOLO: Multi-input Transformer for Age and Gender Estimation. arXiv preprint arXiv:2307.04616 (2023)

11. Gollapudi, S., Gollapudi, S.: OpenCV with python. Learn Computer Vision Using OpenCV: With Deep Learning CNNs and RNNs 31-50 (2019)

12. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)

13. Chen, X., Hsieh, C.-J., Gong, B.: When vision transformers outperform resnets without pre-training or strong data augmentations. arXiv preprint arXiv:2106.01548 (2021)

14. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5810-5818. (2017)

15. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. IEEE Transactions on information forensics and security 9, 2170-2179 (2014)

16. Lanitis, A., Taylor, C.J., Cootes, T.F.: Toward automatic simulation of aging effects on face images. IEEE Transactions on pattern Analysis and machine Intelligence 24, 442-455 (2002)

17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. (2016)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. (2016)