Human Facial Age Group Recognizer using Assisted Bottleneck Transformer Encoder

Adri Priadana[0000−0002−1553−7631], Duy-Linh Nguyen[0000−0001−6184−4133], Xuan-Thuy Vo[0000−0002−7411−0697], and Kang-Hyun Jo*[0000−0002−4937−7082]

Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, South Korea
priadana3202@mail.ulsan.ac.kr, ndlinh301@mail.ulsan.ac.kr, xthuy@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

Abstract. Recognizing age from facial images has attracted considerable attention because of its wide array of applications and practical utilities. These include support for advertising platforms, access control, forensic objectives, and video surveillance. Efficient facial age recognition for these varied purposes is essential, necessitating smooth operation on low-cost devices or, at the very least, on a CPU to minimize implementation costs. This work proposes a lightweight CNN architecture efficiently integrated with a transformer encoder to perform facial age group recognition. An assisted bottleneck transformer encoder (ABTE) is introduced to enhance the feature extractor, generating only a few parameters and requiring low computation. As a result, the proposed architecture can achieve competitive performance on the two benchmark datasets, UTKFace and FG-NET. Moreover, this recognizer can attain real-time speed at 147 and 136 frames per second (FPS) with a single and double utilization of the ABTE, respectively, while maintaining its performance.

Keywords: Age Group Recognition · Assisted Bottleneck Transformer · Convolutional Neural Network (CNN) · Facial Age Recognition · Transformer Encoder.

1 Introduction

Age estimation from facial images has garnered significant interest due to its broad range of applications and practical uses, including support for advertising platforms [22,16], access control [7], forensic applications and video surveillance [2]. In advertising applications, it can assist platforms in audience segmentation and delivering relevant ads and products. For example, in some countries, vending machines can suggest beverages like alcohol or tobacco based on facial age estimation, ensuring compliance with age restrictions for specific items. In forensic applications, it can be used to determine victim or criminal profiles. In the context of surveillance and access control, it can be employed to restrict access to specific areas for individuals of particular age groups. Age recognition involves automatically predicting a person's exact age [26] or categorizing them based on face into age groups [13,16] such as child, teen, adult, and old.

As a popular deep learning technique, Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in age estimation based on facial images. Many studies [1,5,11,20,25] have utilized and proposed deeper CNN architectures to enhance their performance. However, it frequently yields architectures with a significant parameter count, potentially causing operational inefficiency. This limitation can hinder implementation on platforms or machines that utilize low-cost or CPU devices. Hence, the need for efficient architectures with reduced computational demands is evident.

Recently, several efforts have been directed towards designing CNN architectures that are more efficient [18,16] and faster [15], generating few parameters and low operation for enhanced overall efficiency and speed. Moreover, the Vision Transformer (ViT) technique [6] and its variant, inspired by the Transformer architecture [24], initially designed for machine translation tasks, has become dominant and proven to offer high classification performance in computer vision tasks. However, Transformer-based architectures often prioritize accuracy over computational efficiency, which is critical for operation on resource-constrained devices, such as CPUs or mobile platforms. By combining CNN architectures with the Transformer encoder in an efficient manner, it is possible to create models with fewer parameters and reduced computational demands while maintaining or even improving performance. Therefore, it can be satisfactorily performed on a lower-cost device and contribute more to procurement cost reduction.

This work proposes a lightweight CNN architecture integrated efficiently with a transformer encoder to perform a facial age group recognition task. A novel assisted bottleneck transformer encoder, improved from [14], is introduced to enhance the feature extractor used in the recognizer. It generates few parameters and low computation. As a result, the age group recognizer can operate more efficiently and rapidly when identifying age groups based on facial features. To summarize, the notable contributions of the present study include the following:

1. A lightweight CNN architecture integrated with a transformer encoder to perform age group recognition based on facial features. It demonstrates highly competitive performance on UTKFace [29] and FG-NET [10] benchmark datasets.
2. A novel assisted bottleneck transformer encoder (ABTE), inspired by [14], is offered as a strategy to capture spatial relationship representations within the feature maps. The enhancement significantly improves the quality of feature maps, leading to enhanced recognition performance.
3. A facial age group recognizer capable of swift operation on a CPU device. It can achieve real-time performance at 147 and 136 frames per second (FPS) with one and two times assisted bottleneck transformer encoder, respectively.

2 Related Work

Due to the exceptional capabilities of Convolutional Neural Networks (CNNs), the majority of studies in recent years have adopted this approach for age recognition based on human faces. For example, Li et al. [11] introduced BridgeNet, which incorporates local regressors in learning continuity-aware weights for age recognition from facial images. Badr et al. [1] adopted ResNet-34 as a foundation to develop a system called landmark ratios with task importance (LRTI) for age estimation. Another researcher [5] presented a feature constraint reinforcement network (FCRN) for leveraging the influence of gender constraints on age estimation. Meanwhile, Shin et al. [20] utilized the VGG architecture as an encoder, introducing a novel moving window regression algorithm designed to estimate facial age precisely. Wang et al. [25] adopted the ResNet34 and proposed a meta-set learning (MSL) approach for exploiting the unfairness of face-aging datasets.

There has been a trend in developing lightweight CNN architectures to consider efficient computation applied for mobile or CPU-based devices. Savchenko [18] employed MobileNet for estimating facial age in mobile applications, producing a model with 3.5 million parameters. In a different study [16], an efficient CNN architecture was introduced, boasting a mere four hundred thousand parameters. This architecture comprises two branches of feature extractors boosted by an innovative attention mechanism. In the most recent advancement [15], a novel efficient CNN architecture is presented, integrating a lightweight backbone featuring a combination of different mini-feature map levels stimulated by a slight attention block. This network can perform real-time facial age recognition on CPU devices.

3 The Proposed Method

The proposed CNN model is designed with proficient feature extraction and classification phases, as illustrated in Fig. 1. This architecture stands out for its efficiency, boasting a mere 446,468 parameters and approximately 24 million floating-point operations (MFLOPs).

3.1 The Feature Extraction Module

Following a common optimal design approach, the proposed seamless feature extraction strategically utilizes a shallow convolution layer, each employing a 3×3 filter size for precision and effectiveness. It initiates with 16 channels, followed by increments to 32 and, ultimately, to 64. This deliberate design intention aims to minimize the number of parameters and computational burden within the architecture. Additionally, the architecture utilizes batch normalization (BN) after each convolution operation, followed by sigmoid linear units (SiLU) activation, to address gradient-related issues. It incorporates three max-pooling operations as pivotal for downsampling the feature map effectively. These operations employ two 3×3 and one 2×2 kernel sizes, each with strides set at 2. The careful selection of these parameters facilitates a systematic reduction in the spatial dimensions of the feature map, contributing to enhanced efficiency in subsequent processing stages. The downsampling mechanism is strategically

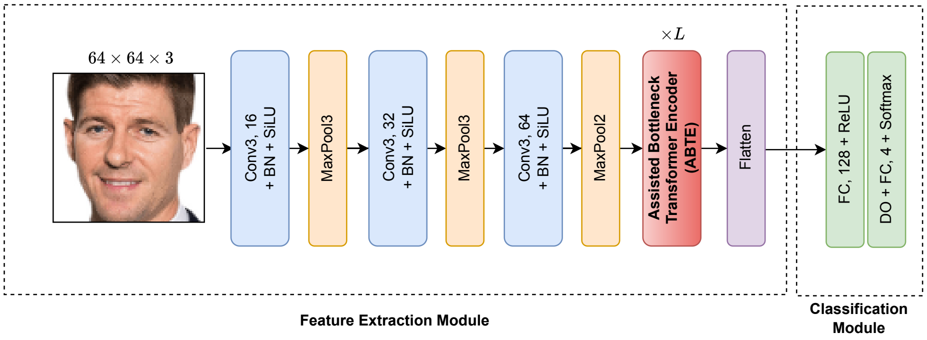


Fig. 1. The proposed lightweight CNN architecture integrated efficiently with a assisted bottleneck transformer encoder to perform facial age group recognition. Conv3 indicates 3 × 3 convolution operation with stride 1. MaxPool3 and Maxpool2 denote 3 × 3 and 2 × 2 max pooling operations with stride 2, respectively. BN, DO, and FC are batch normalization, dropout, and fully connected layers.

positioned to optimize the overall computational load, ensuring the architecture's responsiveness to real-time demands. It is important to note that opting for fewer convolution layers can result in a shallower network, which impacts its performance. In response, we introduce an assisted bottleneck transformer encoder (ABTE) to capture spatial relationship representations within the feature maps by applying self-attention as the core operation, enhancing recognition performance. The proposed architecture situates this encoder between the final max-pooling and flattening operation in the layer sequence.

3.2 The Assisted Bottleneck Transformer Encoder (ABTE)

Nowadays, Vision Transformer (ViT) [6] with self-attention has shown impressive performance in image classification tasks. The self-attention mechanism allows the model to capture relationships between different parts of the input image using a global context. It makes the architecture particularly effective for tasks that require understanding long-range relationships within an image. ViT has achieved state-of-the-art performance on various image classification benchmarks. It attains competitive accuracy with fewer parameters. However, the computational efficiency of ViT still becomes a concern because the self-attention mechanism has a quadratic complexity regarding the input sequence length. Many researchers have proposed various techniques to address these issues. This work proposes an assisted bottleneck transformer encoder (ABTE). It consists of an efficient transformer encoder, assisted by an enhanced efficient convolution module shown in Fig. 2. Following the efficient transformer encoder in [14], a bottleneck transformer encoder (BTE) is applied to enhance the computational efficiency of the encoder. BTE employs a reduction channel denoted as r and a multi-head mechanism with $NumHead$ to generate a more streamlined input tensor of dimensions $H \times W \times ((\frac{C}{NumHead})/r)$ before transforming

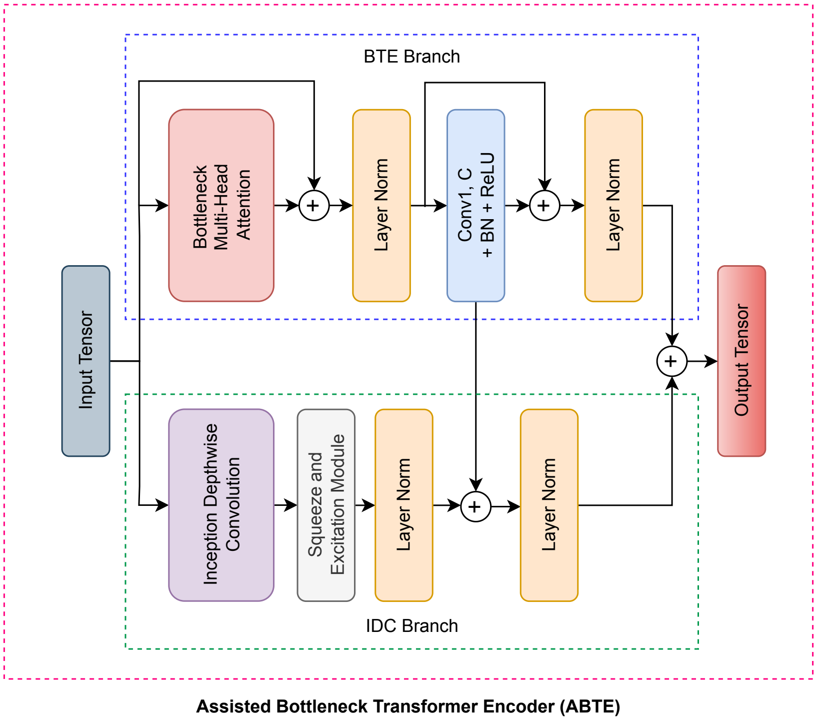

BTE Branch
Bottleneck Multi-Head Attention
Layer Norm
Conv1, C + BN + ReLU
Layer Norm
Input Tensor
Output Tensor
Inception Depthwise Convolution
Squeeze and Excitation Module
Layer Norm
Layer Norm
IDC Branch
Assisted Bottleneck Transformer Encoder (ABTE)

Fig. 2. The proposed assisted transformer encoder consists of an efficient transformer encoder, assisted by an efficient inception depthwise convolution module.

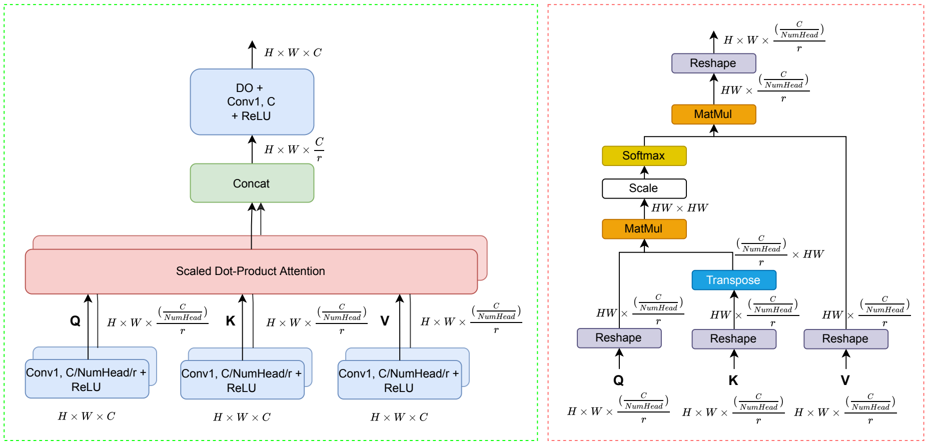
it into a query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$). The implementation of scaled dot-product attention, as depicted in Fig. 3, is similar to the structure of the original transformer encoder and is defined by the following definition:

$$\text{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

where T is a transpose matrix operation and d_k indicates a scaling factor to control the softmax temperature. Improved from BTE, we apply an inception depthwise convolution (IDC) module from InceptionNeXt [28] in parallel structure to assist the BTE. This module decomposes multi-kernel depthwise convolution into four parallel branches along the channel dimension. It splits the input feature map $\mathbf{X}$ into four elements $[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}]$ based on channel axes and applies small square kernels (3×3), two orthogonal band kernels (11×1 and 11×11), and an identity mapping, respectively, defined as follows:

$$\text{IDC}(\mathbf{X}) = \text{Concat}[\text{DW}_{3\times3}(\mathbf{X_1}), \text{DW}_{1\times11}(\mathbf{X_2}), \text{DW}_{11\times1}(\mathbf{X_3}), \mathbf{X_4}], \tag{2}$$

where Concat and $\mathrm{DW_{m\times n}}$ indicate concatenation and depthwise convolution operations with $m \times n$ kernel size, respectively. IDC module function to preserve performance by efficiently applying multi-scale and large-kernel-based convolution. We also utilize squeeze-and-excitation (SE) [9] module and layer normalization to enhance and normalize the feature maps resulting from the IDC module. Fig. 4 shows the IDC and SE modules in more detail. In this work, ABTE performs interaction or connection between BTE and IDC branches to share information by adding the feature map resulting from the convolution operation in the BTE branch to the IDC branch using an element-wise addition operation. The proposed encoder also uses this mechanism to allocate a more significant portion to BTE in extracting information. Based on the model analysis results described in the ablation study subsection in the next section, BTE contributes more performance than IDC when performing individually. Moreover, this ABTE combines the BTE and IDC branches by applying an element-wise addition operation in the last layer.


$H \times W \times C$
DO + Conv1, C + ReLU
$H \times W \times \frac{C}{r}$
Concat
Scaled Dot-Product Attention
Q
K
V
$H \times W \times \frac{(\frac{C}{NumHead})}{r}$
Conv1, C/NumHead/r + ReLU
$H \times W \times C$
Reshape
$HW \times \frac{(\frac{C}{NumHead})}{r}$
MatMul
Softmax
Scale
$HW \times HW$
Transpose
$\frac{(\frac{C}{NumHead})}{r} \times HW$

Bottleneck Multi-Head Attention

Scaled Dot-Product Attention

Fig. 3. The detail design of the bottleneck multi-head attention [14].

3.3 The Classification Module

In the final phase, the features of the instance face coming from the backbone phase are inputted into the classification module to calculate the probability for individual group classes. This component assists in determining whether the instance input belongs to which class individual. It consists of two multi-layer perceptron layers, following the classification module from [15].

4 Implementation Settings

Following the previous setting [15], the proposed architecture undergoes training on UTKFace and FG-NET datasets with an initial learning rate of 1×10^{-3}

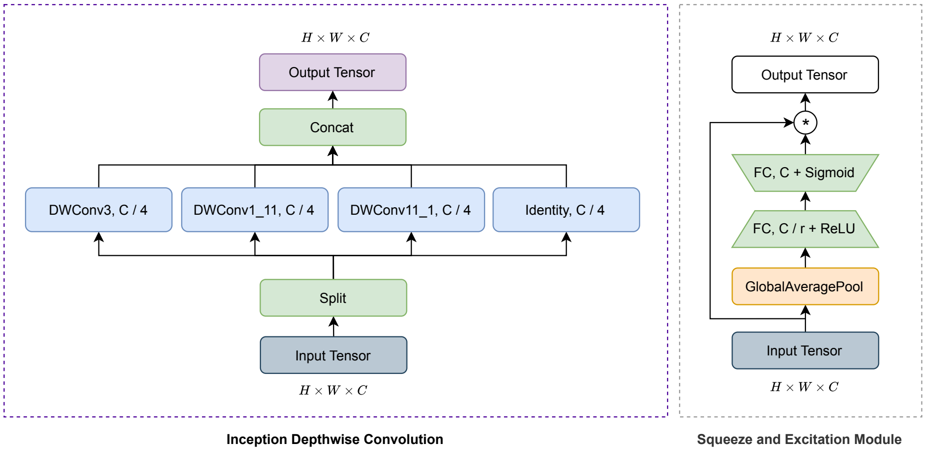


Fig. 4. The inception depthwise convolution (IDC) and squeeze and excitation (SE) modules.

with a batch size of 256 trained over 300 epochs and Adam optimizer. In this configuration, the reduce learning rate mechanism is implemented to modulate the learning rate based on changes in validation accuracy. The rate is decreased by a factor of 0.75 after 20 epochs of stagnant accuracy, contributing to adaptive learning in the training process. It utilizes an Nvidia GeForce GTX 1080Ti featuring GPUs with 11GB of memory through the Tensorflow and Keras framework. An Intel Core i7-9750H CPU running at 2.6 GHz with 20GB of RAM is employed to evaluate the FPS for both the proposed architecture and the recognizer.

5 Experiments and Results

5.1 Evaluation on Datasets

UTKFace. This dataset is a widely utilized resource in the field of computer vision, particularly for research on age and gender estimation. It comprises 23,708 facial images annotated with valuable information such as age, gender, and ethnicity. Notably, the age range covered by the dataset spans from 0 to 116 years, and it incorporates diverse images with variations in pose, illumination, expression, and other factors. This work applies two configurations for the dataset as an evaluation. Following the prior studies [4,3], this dataset is divided into training (80%) and testing (20%) sets for the first configuration, denoted as Setting I. The evaluation of the offered architecture's performance involves the calculation of the mean absolute error (MAE) on the testing set within this configuration. The proposed architecture, comprising only around 450,000 parameters, achieves the second-best performance with the mean absolute error

Table 1. The results of the assessment on the UTKFace dataset under Setting I.

Architectures	**Params (M)**	**Mean Absolute Error ↓**
CORAL [4]	21.11	5.47
Savchenko [18]	3.50	5.44
LRTI [1]	21.11	4.55
Berg et al. [3]	23.60	5.14
FCRN [5]	23.60	4.47
2PDG [16]	0.46	4.44
AggerCPU [15]	0.49	4.38
MWR (based on VGG16) [20]	39.79	4.37
MSL (based on ResNet34) [25]	21.11	**4.31**
Proposed ($L = 2$)	**0.45**	**4.37**

Table 2. The results of the assessment on the UTKFace dataset under Setting II.

Architectures	**Params (M)**	**Validation Accuracy ↑ (%)**
ResNet50 [8]	23.60	88.43
InceptionNeXt-N16 [28]	0.36	90.08
2PDG [16]	0.46	90.12
VGG16 [21]	39.79	90.34
InceptionNeXt-N24 [28]	0.80	90.81
AggerCPU [15]	0.46	90.90
Proposed ($L = 2$)	**0.45**	**91.33**

(MAE) of 4.37, as shown in Table 1. The result is marginally lower by only 0.06 compared to the top-performing model [25]. The proposed architecture proves significantly more efficient than [25] considering the number of parameters.

Similarly, following the methodology of a prior study [15], we divided the dataset into training (90%) and testing (10%) sets for the second configuration, identified as Setting II. The class target comprises four age groups: children, teens, adults, and old. Validation accuracy (VA) is utilized in evaluating the proposed architecture in this setting. The offered architecture achieves the VA of 91.33%, surpassing the state-of-the-art, as shown in Table 2. This experiment also compares our proposed model with InceptionNeXt [28] model, -N16 and -N24, which means applied 16 and 24 as initial embedding dimensions, respectively, of InceptionNeXt T model configuration, to generate the InceptionNeXt [28] variant that has a comparable number of parameters with our proposed model.

FG-NET. This dataset comprises 1,002 facial images collected from 82 subjects, featuring variations in pose, expression, and illumination. Following established settings [19,12], the dataset employs k-fold cross-validation and leave-one-

Table 3. The results of the assessment on the FG-NET dataset.

Architectures	**Params (M)**	**Mean Absolute Error ↓**
DRF based on VGG16 [19]	14.00	3.41
DAG-VGG16 [23]	24.00	3.08
ADPF [26]	14.00	2.86
2PDG [16]	0.46	2.75
MSFCL [27]	15.00	2.71
AggerCPU [15]	0.49	2.71
BridgeNet [11]	120.00	**2.56**
MWR based on VGG16 [20]	40.00	**2.23**
Proposed ($L = 2$)	**0.45**	**2.67**

person-out (LOPO) methodologies. In each fold, facial images from one subject are reserved for testing, while the images of the remaining subjects are used for training. This process is repeated 82 times, with each subject applied as a training set, corresponding to the 82 subjects in the dataset. Given the diverse distribution of instances among individuals in the dataset, the number of instances for both training and testing sets exhibits variability across each fold. It is important to note that this evaluation process computes results based on average values using the mean absolute error (MAE) metric. The proposed architecture achieves the MAE of 2,67, attaining the third-best performance, deviating by 0.44 and 0.11 from the best [20] and second-best [11], respectively, as shown in Table 3. However, the parameters of the proposed CNN model are significantly lower than both.

5.2 Model Analysis

This section investigates the contribution of each component of the proposed module to the recognition performance on the UTKFace dataset. Firstly, we perform an ablation study by removing each module from the proposed model and then comparing its performance to reveal the influence of the existence of each module. Secondly, channel reduction analysis examines the optimal channel reduction value on the BTE. Lastly, we analyzed how many times the ABTE (the number of L) should be applied to produce the best performance.

Ablation Study. Table 4 shows the reported results of the ablation study based on validation accuracy metrics. The report shows that using IDC and BTE modules individually can escalate performance based on accuracy by 0.35% and 0.48%, respectively, from the baseline based on accuracy. Combining IDC and BTE modules can upgrade performance by 0.74% from the baseline. Moreover, combining IDC and BTE modules with a connection can enhance performance by 1.08% from the baseline. These results confirm that the IDC module can assist the BTE in providing higher performance.

Table 4. Ablation study of the proposed architecture with $L = 2$ on the UTKFace dataset under Setting II.

Baseline	**IDC Branch**	**BTE Branch**	**Connection between IDC and BTE**	**MFLOPs**	**Params**	**Validation Accuracy (%)**
✓				22.15	426,084	90.25
✓	✓			22.33	428,452	90.60
✓		✓		23.95	443,844	90.73
✓	✓	✓		24.13	446,212	90.99
✓	✓	✓	✓	**24.18**	**446,468**	**91.33**

Table 5. Channel reduction analysis of the proposed ABTE with $L = 2$ on the UTKFace dataset under Setting II.

Value of Reduction r	**MFLOPs**	**Params**	**Validation Accuracy (%)**
1	27.38	471,332	90.94
2	25.15	454,756	91.12
4	**24.18**	**446,468**	**91.33**
8	23.74	442,324	90.86

Channel Reduction Analysis. This study explores the impact of different channel reduction values on the performance of the proposed ABTE in facial recognition. The findings, as presented in Table 5, suggest that channel reduction values of one or eight do not yield significant performance improvements. The optimal recognition performance is achieved by the proposed ABTE when employing a channel reduction value of four. In this configuration, the model attains the highest validation accuracy, reaching 91.33%, while maintaining a moderate parameter count of 446,468 and a computational load of 24.18 MFLOPs.

Number of Transformer Encoder Analysis. This part analyzes the most effective number of L (how many times applying ABTE) on the proposed model regarding recognition performance. Table 6 reveals that using ABTE one time can make the model run faster. However, the proposed model with two times ABTE achieves the highest validation accuracy in this study. This setting can deliver a validation accuracy of 91.33% with 446,468 parameters and 24.18 MFLOPs at a sufficient speed based on FPS. Age (FPS) denotes the speed of age group recognition, and Face + Age (FPS) indicates the speed of age group recognition integrated with face detection [17].

5.3 Runtime Efficiency

The practical implementation prioritizes a recognizer capable of real-time performance on cost-effective devices, ideally on a CPU setup, to reduce expenses during system procurement. The offered architecture, featuring two times of

Table 6. Number of L analysis on the UTKFace dataset under Setting II.

Number of L	**MFLOPs**	**Params**	**Validation Accuracy (%)**	**Age (FPS)**	**Face + Age (FPS)**
1	23.17	436,276	91.07	**335.99**	**147.53**
2	24.18	446,468	**91.33**	285.67	135.48
3	25.20	456,660	90.85	250.43	127.52

Table 7. The efficiency of runtime, as measured on the UTKFace dataset with the same CPU configuration, is specifically evaluated under Setting II.

Architectures	**MFLOPs**	**Params**	**Validation Accuracy (%)**	**Age (FPS)**	**Face + Age (FPS)**
VGG16 [21]	2,290	39,782,722	90.34	42.40	36.28
ResNet50 [8]	633	23,595,908	88.43	54.21	44.54
InceptionNeXt-N24 [28]	1,391	796,564	90.81	57.44	46.87
InceptionNeXt-N16 [28]	625	359,012	90.08	89.09	65.99
Proposed ($L = 2$)	**24**	**446,468**	**91.33**	**285.67**	**135.48**
AggerCPU [15]	41	486,822	90.90	330.66	144.49
Proposed ($L = 1$)	**23**	**436,276**	**91.07**	**335.99**	**147.53**

ABTE ($L = 2$), demonstrates real-time efficiency on a CPU with a modest parameter count of 446,468 and a computational load of 24.18 MFLOPs. It excels in classification tasks, achieving a speed of 286 and 135 frames per second for age group recognition and integrated with face detection [17] (Face + Age), respectively, as detailed in Table 7.

This work also presents a recognizer that performs ABTE only once ($L = 1$) utilizing 436,276 parameters and 23.17 MFLOPs to provide faster recognition with a performance that still surpasses the current state-of-the-art models. The proposed recognizer, leveraging a single ABTE operation, emerges as the fastest among competitors, achieving a remarkable speed of 336 frames per second for age group recognition (Age) and 148 frames per second for age group recognition integrated with face detection [17] (Face + Age). The recognition outcomes of the proposed model are illustrated in Fig. 5, where green, yellow, blue, and red bounding boxes signify the faces of children, teens, adults, and old, respectively.

6 Conclusion

Addressing the need for improved feature extraction in age group recognition from human faces, this work introduces the concept of ABTE. By incorporating a bottleneck mechanism and employing inception depthwise convolution (IDC) in parallel structure, the proposed encoder efficiently enhances the transformer encoder's capabilities while maintaining a minimal parameter count and low computational requirements. Demonstrating competitive performance on UTKFace and FG-NET datasets, the proposed architecture doubles as a recognizer,

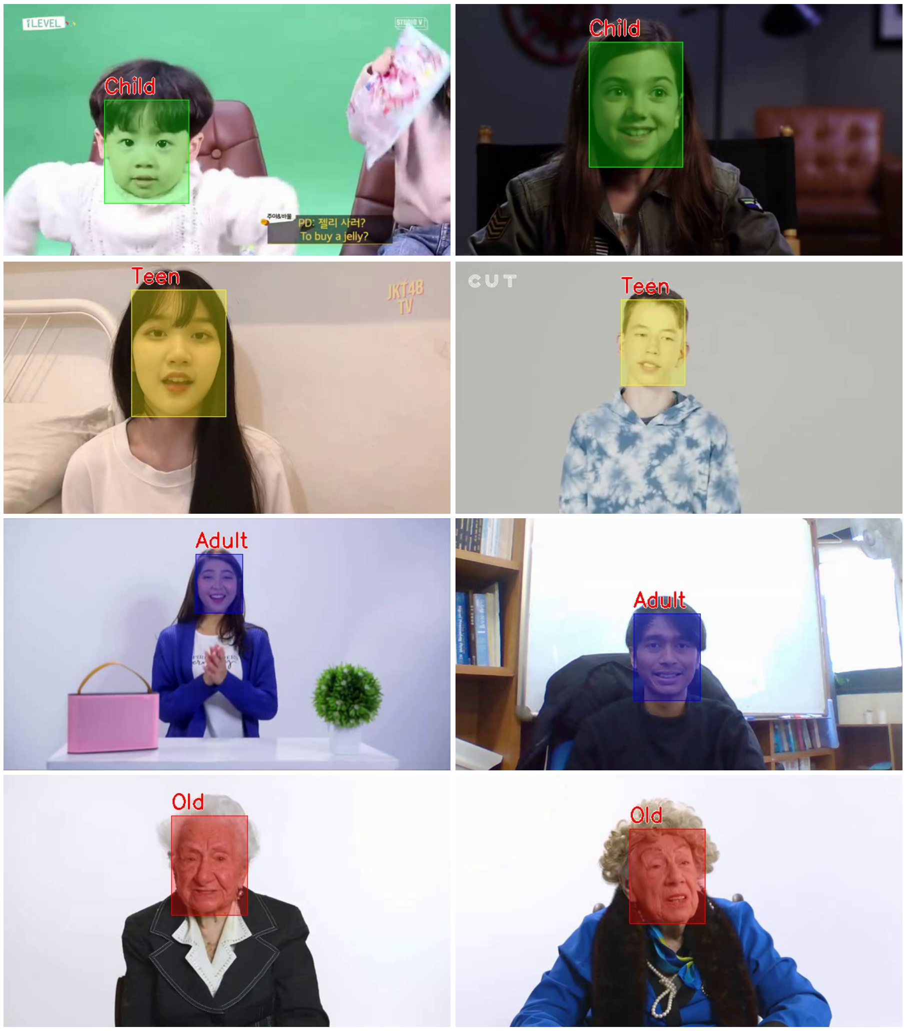


Fig. 5. The example of the recognition results of the proposed recognizer.

achieving real-time speeds of 147 and 136 FPS with a single and double utilization of the assisted bottleneck transformer encoder, respectively. As part of future endeavors, the proposed facial age group recognizer will be extended to operate on more cost-effective devices, further supporting applications in Robot Vision.

Acknowledgment

This result was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003).

References

1. Badr, M.M., Elbasiony, R.M., Sarhan, A.M.: Lrti: landmark ratios with task importance toward accurate age estimation using deep neural networks. Neural Computing and Applications **34**(12), 9647–9659 (2022)
2. Becerra-Riera, F., Morales-González, A., Méndez-Vázquez, H.: A survey on facial soft biometrics for video surveillance and forensic applications. Artificial Intelligence Review **52**(2), 1155–1187 (2019)
3. Berg, A., Oskarsson, M., O'Connor, M.: Deep ordinal regression with label diversity. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 2740–2747. IEEE (2021)
4. 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)
5. Chen, G., Peng, J., Wang, L., Yuan, H., Huang, Y.: Feature constraint reinforcement based age estimation. Multimedia Tools and Applications **82**(11), 17033–17054 (2023)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Gupta, S.K., Nain, N.: Single attribute and multi attribute facial gender and age estimation. Multimedia Tools and Applications **82**(1), 1289–1311 (2023)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE (2016)
9. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(8), 2011–2023 (2019)
10. 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**(4), 442–455 (2002)
11. Li, W., Lu, J., Feng, J., Xu, C., Zhou, J., Tian, Q.: Bridgenet: A continuity-aware probabilistic network for age estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1145–1154 (2019)
12. Liu, H., Lu, J., Feng, J., Zhou, J.: Label-sensitive deep metric learning for facial age estimation. IEEE Transactions on Information Forensics and Security **13**(2), 292–305 (2017)
13. Mai, A.T., Nguyen, D.H., Dang, T.T.: Real-time age-group and accurate age prediction with bagging and transfer learning. In: 2021 International Conference on Decision Aid Sciences and Application (DASA). pp. 27–32. IEEE (2021)
14. Priadana, A., Putro, M.D., An, J., Nguyen, D.L., Vo, X.T., Jo, K.H.: Gender recognizer based on human face using cnn and bottleneck transformer encoder. In: 2023 International Workshop on Intelligent Systems (IWIS). pp. 1–6. IEEE (2023)

15. Priadana, A., Putro, M.D., Nguyen, D.L., Vo, X.T., Jo, K.H.: Age group recognizer based on human face supporting smart digital advertising platforms. In: 2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE). pp. 1–7. IEEE (2023)
16. Priadana, A., Putro, M.D., Vo, X.T., Jo, K.H.: An efficient face-based age group detector on a cpu using two perspective convolution with attention modules. In: 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR). pp. 1–6. IEEE (2022)
17. Putro, M.D., Nguyen, D.L., Jo, K.H.: Lightweight convolutional neural network for real-time face detector on cpu supporting interaction of service robot. In: 2020 13th International Conference on Human System Interaction (HSI). pp. 94–99. IEEE (2020)
18. Savchenko, A.V.: Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output convnet. PeerJ Computer Science **5**, e197 (2019)
19. Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., Yuille, A.: Deep differentiable random forests for age estimation. IEEE transactions on pattern analysis and machine intelligence **43**(2), 404–419 (2019)
20. Shin, N.H., Lee, S.H., Kim, C.S.: Moving window regression: A novel approach to ordinal regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18760–18769 (2022)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Suman, S., Urolagin, S.: Age gender and sentiment analysis to select relevant advertisements for a user using cnn. In: Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2021, pp. 543–557. Springer (2022)
23. Taheri, S., Toygar, Ö.: On the use of dag-cnn architecture for age estimation with multi-stage features fusion. Neurocomputing **329**, 300–310 (2019)
24. 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)
25. Wang, C., Li, Z., Mo, X., Tang, X., Liu, H.: Exploiting unfairness with metaset learning for chronological age estimation. IEEE Transactions on Information Forensics and Security (2023)
26. Wang, H., Sanchez, V., Li, C.T.: Improving face-based age estimation with attention-based dynamic patch fusion. IEEE Transactions on Image Processing **31**, 1084–1096 (2022)
27. Xia, M., Zhang, X., Weng, L., Xu, Y., et al.: Multi-stage feature constraints learning for age estimation. IEEE Transactions on Information Forensics and Security **15**, 2417–2428 (2020)
28. Yu, W., Zhou, P., Yan, S., Wang, X.: Inceptionnext: when inception meets convnext. arXiv preprint arXiv:2303.16900 (2023)
29. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4352–4360. IEEE (2017)