Simultaneous Facial Age Group and Gender Recognition using Efficient Local-Global Attention Network for Intelligent Advertising

Adri Priadana[1], Duy-Linh Nguyen[1], Xuan-Thuy Vo[1], Muhamad Dwisnanto Putro[2], Ge Cao[1], and Kanghyun Jo[1]
[1]*Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, Ulsan, Korea*
[2]*Department of Electrical Engineering, Universitas Sam Ratulangi, Manado, Indonesia*
Email: priadana3202@mail.ulsan.ac.kr; ndlinh301@mail.ulsan.ac.kr; xthuy@islab.ulsan.ac.kr;
dwisnantoputro@unsrat.ac.id; caoge@islab.ulsan.ac.kr; acejo@ulsan.ac.kr

***Abstract*—Facial age group and gender recognition have attracted significant interest due to their wide range of applications and practical uses, including aiding advertising platforms in delivering relevant content. Performing efficient simultaneous recognition of facial age group and gender is crucial for these applications, necessitating seamless performance on inexpensive devices to mitigate implementation expenses. This work introduces an Efficient Local-Global Attention Network (ELGA-Net) for facial age group and gender recognition simultaneously. It proposes an Efficient Local-Global Attention (ELGA) block to enhance the quality of feature maps locally while capturing global contextual information by learning relationships between different parts of the feature maps. As a result, the proposed network achieves competitive performance on the UTKFace dataset. Moreover, it attains real-time speed at 113 frames per second (FPS) on an Intel Core i7-9750H CPU 2.6 GHz device when integrated with face detection as an initial process.**

***Index Terms*—Age group and gender recognition, intelligent advertising, local global attention, multi-task network, real-time recognition.**

I. Introduction

The landscape of outdoor and offline advertising has witnessed notable developments. It includes digital technologies integration into traditional methods, such as digital signage [1] and billboards [2] with a focus on dynamic content and interactive displays. These platforms leverage data analytics to meet campaign effectiveness, incorporating artificial intelligence technology for optimizing advertising. Moreover, computer vision technology facilitates real-time audience analytics by performing object recognition, such as facial attributes [3], for dynamic content adjustments. According to those analytic results, the advertising platform can provide more relevant promotional content.

In the advertising industry, age and gender are significant attributes as they provide valuable demographic insights for targeted marketing [4]. Recognizing the age or age group and gender of the target audience enables advertisers to create more relevant and tailored campaigns, including factors like preferences, interests, and buying behaviors. This information helps align products with the needs and interests of specific demographic groups, ensuring that advertising messages resonate effectively. It would result in improved engagement, higher conversion rates, enhanced brand perception, optimized marketing budgets, and a better customer experience.

Age group and gender recognition can be performed through a face, utilizing computer vision and deep learning models to analyze facial features and estimate a person's age group and gender from images. These systems employ deep convolutional neural networks (DCNNs) and diverse training datasets to learn patterns within facial data. In recent studies addressing age-related tasks in computer vision, researchers have proposed innovative approaches to overcome challenges inherent in facial age recognition. A meta-set learning approach [5] was introduced that leverages unfairness in facial age datasets to achieve unbiased age classification in diverse conditions. Another work [4] focused on developing an efficient DCNN architecture incorporating a residual mini multi-level and deep lite attention module for performing age group recognition. For gender recognition, DCNN architectures using transfer learning approaches, the research in [6] employed DCNN architectures using transfer learning approaches. A bottleneck transformer encoder [7] was initiated to increase recognition performance by executing global context learning efficiently.

Performing age group and gender recognition individually using different independent networks can potentially lead to increased computational complexity and parameters. This scenario may require additional computational power and memory, thus resulting in higher hardware requirements. Even age or gender recognition requires face detection as an initial process [8]. Therefore, adopting a unified network in simultaneously executing age group and gender recognition can offer more resource efficiency. Liao et al. [9] designed an effective multi-task architecture that learns gender and age together, leveraging the dependency between these attributes to enhance recognition accuracy. It introduces a random forest method for extracting robust multi-instance and multi-scale features to mitigate the impact of intra-subject distortions. Another study in [10] utilized DCNN architecture with a pre-trained mechanism for gender and age estimation using a multi-tasking approach, demonstrating promising performance. Unfortunately, the DCNN generates a lot of parameters and operations, leading to operating slowly, especially on low-cost

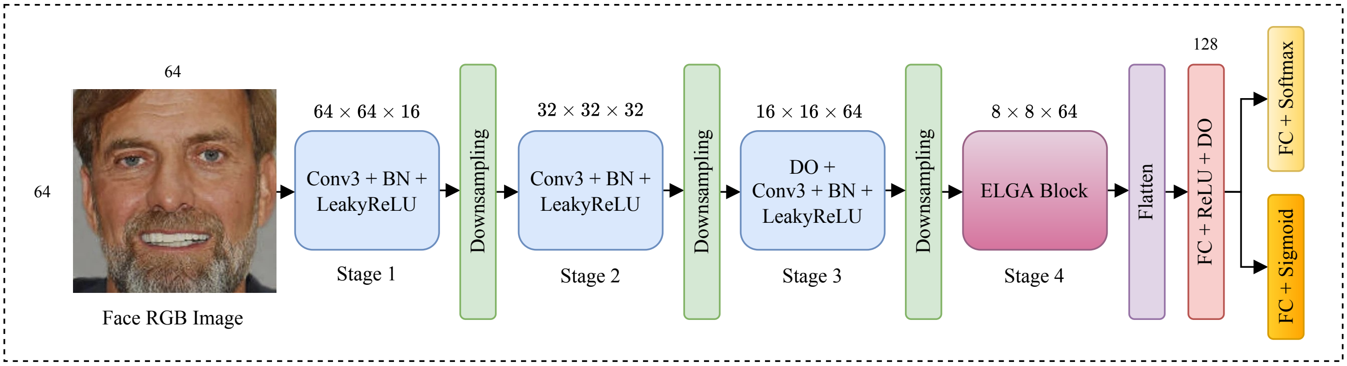

Fig. 1. The efficient local-global attention network (ELGA-Net) for simultaneous facial age group and gender recognition. Conv3, BN, DO, FC, and Downsampling indicate 3×3 convolution with strides 1, batch normalization, dropout, fully connected, and 2×2 max-pooling with strides 2, respectively. ReLU, LeakyReLU, Sigmoid, and Softmax are activation functions. The hyperparameters in this network, such as the number of channels, are empirically set.

or CPU devices used on an advertising platform.

In light of parameters, computation, and speed, this work proposes an Efficient Local-Global Attention Network (ELGA-Net) for simultaneously performing facial age group and gender recognition. It introduces an Efficient Local-Global Attention (ELGA) block to enhance the quality of feature maps locally while capturing global contextual information by learning relationships between different parts of the feature maps. This network generates a few parameters and requires low computational resources, making it ideal to implement on a low-cost or CPU device. This work delineates its contributions as follows:

1) An Efficient Local-Global Attention Network (ELGA-Net) to perform simultaneous facial age group and gender recognition. It shows promising performance on the UTKFace [11] benchmark dataset.
2) A novel Efficient Local-Global Attention (ELGA) block is offered to capture local (supported by attention modules) and global information within the feature maps. It can enhance the quality of feature maps, guiding to improve recognition performance.
3) A simultaneous facial age group and gender recognizer designed for rapid execution on a CPU device. It demonstrates real-time performance, achieving 113 frames per second (FPS) when integrated with a face detector.

II. Proposed Architecture

The efficient network presented in this study consists of four stages followed by a multi-task classification module, as illustrated in Fig. 1. This network incorporates efficient local-global attention block located at the fourth stage. The proposed network generates only 552,029 parameters and 24.35 MFLOPs.

A. The Efficient Local-Global Attention (ELGA) Block

Convolutional Neural Networks (CNNs) have long been proven capable of capturing local patterns in images [12], owing to their local receptive fields. On the other hand, Vision Transformers (ViTs) can capture global dependencies and relationships across different parts of the image by performing a self-attention technique [13]. CNNs are adept at feature extraction, while Transformers can understand the broader context. This work proposes a novel Efficient Local-Global Attention (ELGA) block to leverage the strengths of both CNNs and Transformers. It consists of an efficient CNN block, integrated by an efficient transformer encoder shown in Fig. 2. ELGA conveys more significant efficiency via a channel-splitting mechanism to adopt parallel convolution and self-attention paths as local and global information extractor modules. Technically, given the input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, it is split into $\mathbf{X_l} \in \mathbb{R}^{H \times W \times \frac{C}{2}}$ and $\mathbf{X_g} \in \mathbb{R}^{H \times W \times \frac{C}{2}}$. Then it performs local (supported by attention modules) and global modules on $\mathbf{X_l}$ and $\mathbf{X_g}$, respectively, defined as follows:

$$\begin{aligned}\mathrm{ELGA}(\mathbf{X}) = \mathrm{LN}(\mathbf{X} + \mathrm{Concat}[&\mathrm{AM}(\mathrm{Local}(\mathbf{X_l})),\\ &\mathrm{Global}(\mathbf{X_g})]),\end{aligned} \tag{1}$$

where Concat and LN are concatenation and layer normalization operations. AM is the attention strategy, applying channel and spatial attention modules. Inspired by [14], this work proposed multi-scale and multi-band-level based depthwise convolution to capture the local information. Given an input feature map $\mathbf{X_l} \in \mathbb{R}^{H \times W \times \frac{C}{2}}$, it is split the input feature map $\mathbf{X_l}$ into $\mathbf{X_{l1}} \in \mathbb{R}^{H \times W \times \frac{C}{4}}$ and $\mathbf{X_{l2}} \in \mathbb{R}^{H \times W \times \frac{C}{4}}$ based on channel axes. Then, it applies a small square kernel (3×3) and a sequence of two orthogonal band kernels (11×1 and 1×11), respectively. It also uses channel and spatial attention modules that apply global average pooling and average pooling across channels, respectively, with sigmoid activation to enhance the feature map quality. The overall local extractor module, enhanced by attention modules, is described as follows:

$$\begin{aligned}\mathrm{Local}(\mathbf{X_l}) = \mathrm{Concat}[&\mathrm{DW}_{3\times3}(\mathbf{X_{l1}}),\\ &\mathrm{DW}_{1\times11}(\mathrm{DW}_{11\times1}(\mathbf{X_{l2}}))],\end{aligned} \tag{2}$$

$$\mathrm{AM}(\mathbf{X_l'}) = \mathrm{SA}(\mathrm{CA}(\mathbf{X_l'})) \tag{3}$$

where $\mathrm{DW}_{m\times n}$ indicates depthwise convolution operations with $m \times n$ kernel size. CA and SA denote channel and spatial attention operations, respectively.

Following the success of the efficient transformer encoder on [7], ELGA utilized the Bottleneck Transformer Encoder

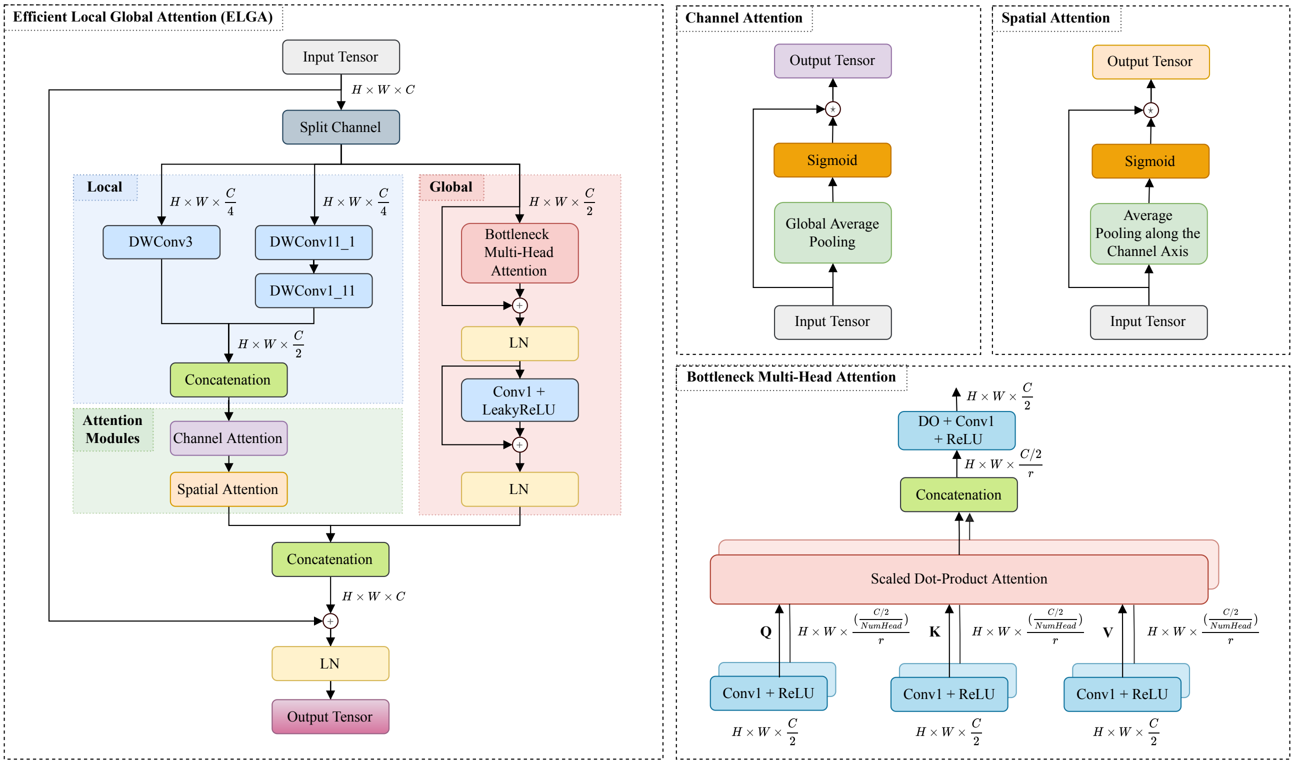
Efficient Local Global Attention (ELGA)
Input Tensor
$H \times W \times C$
Split Channel
Local
$H \times W \times \frac{C}{4}$
$H \times W \times \frac{C}{4}$
DWConv3
DWConv11_1
DWConv1_11
$H \times W \times \frac{C}{2}$
Concatenation
Attention Modules
Channel Attention
Spatial Attention
Global
$H \times W \times \frac{C}{2}$
Bottleneck Multi-Head Attention
LN
Conv1 + LeakyReLU
LN
Concatenation
$H \times W \times C$
LN
Output Tensor
Channel Attention
Output Tensor
Sigmoid
Global Average Pooling
Input Tensor
Spatial Attention
Output Tensor
Sigmoid
Average Pooling along the Channel Axis
Input Tensor
Bottleneck Multi-Head Attention
$H \times W \times \frac{C}{2}$
DO + Conv1 + ReLU
$H \times W \times \frac{C/2}{r}$
Concatenation
Scaled Dot-Product Attention
Q
$H \times W \times \frac{(\frac{C/2}{NumHead})}{r}$
K
$H \times W \times \frac{(\frac{C/2}{NumHead})}{r}$
V
$H \times W \times \frac{(\frac{C/2}{NumHead})}{r}$
Conv1 + ReLU
Conv1 + ReLU
Conv1 + ReLU
$H \times W \times \frac{C}{2}$
$H \times W \times \frac{C}{2}$
$H \times W \times \frac{C}{2}$

Fig. 2. The efficient local-global attention (ELGA) block. Conv, DWConv, DO, and LN indicate convolution, depthwise convolution, dropout, and layer normalization, respectively. ReLU, Leaky ReLU, and Sigmoid are activation function.

(BTE) as a global extractor module to capture global context and relationships between different parts of the feature maps. Given an input feature map $\mathbf{X_g} \in \mathbb{R}^{H\times W\times \frac{C}{2}}$, it computes the feature map into a query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) by involving a 1×1 convolution operation with a reduction channel r and multi-head mechanism with $NumHead$, followed by Rectified Linear Unit (ReLU) to generate a skinnier $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ matrix with shape $H \times W \times ((\frac{C/2}{NumHead})/r)$. Then, we perform these matrices using Scaled Dot-Product Attention (SDPA), which is formulated as follows:

$$\mathrm{SDPA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{4}$$

where T is a transpose matrix operation and d_k is a scaling factor to control the softmax temperature. Subsequently, a concatenation operation combines the outputs from all attention heads, which is followed by a 1×1 convolution operation, incorporating dropout (DO) and ReLU activation layers to restore the channel dimensions to match the input tensor $\mathbf{X_g} \in \mathbb{R}^{H\times W\times \frac{C}{2}}$. Further, a linear projection is executed using a 1×1 convolution operation with LeakyReLU activation, integrated with a residual connection and layer normalization around both the bottleneck multi-head attention and the final convolution layer of the global module.

B. The Overall Network

As shown in Fig. 1, the proposed efficient network comprises four stages following [7]. The initial three stages involve a 3×3 convolution layer with strides of one, followed by batch normalization and LeakyReLU activation. These convolution operations increase the number of channels from 16 to 64. In the third stage, dropout is applied before the convolution operation to address potential overfitting issues. It employs a 2×2 max-pooling operation with strides of two after each stage to facilitate downsampling. Moving to the fourth stage, we introduce the proposed ELGA block, followed by a flattening operation to generate a vector embedding. Subsequently, the multi-task classification module is performed in the last part of the network, consisting of a shared fully connected layer with ReLU activation and dropout mechanism followed by two separate branches or heads for each classification task. These branches have their own fully connected layers customized for the particular task, where, in this work, age group and gender recognition tasks involve utilizing softmax and sigmoid activations, respectively.

III. Implementation Setup

Building upon the methodology established in a previous study [4], [7], we train the proposed architecture on the UTK-Face, FG-NET, and LFW datasets for simultaneous facial age group and gender recognition, as well as individual age and gender recognition, respectively. To perform age and gender recognition separately on the FG-NET and LFW datasets, we remove the final gender or age classification branch from the proposed network, respectively. The training employs an initial learning rate of 10^{-3} with a batch size of 256 over 300

TABLE I
SIMULTANEOUS AGE GROUP AND GENDER RECOGNITION EVALUATION RESULTS ON UTKFACE DATASET.

Networks	**Params**	**MFLOPs**	**Age Accuracy (%) ↑**	**Gender Accuracy (%) ↑**	**Average Accuracy (%) ↑**
InceptionNeXt_N16 [14]	695,845	82.43	86.38	88.54	87.46
GhostNet_N10 [15]	665,765	15.43	86.62	89.16	87.89
FasterNet_N16 [16]	623,621	35.61	89.19	89.70	89.45
VAN_N16 [17]	617,381	76.69	88.74	90.39	89.57
AggerCPU [4]	486,951	40.58	89.77	91.79	90.78
GenderBTE [7]	555,645	24.80	90.26	**92.11**	91.19
Proposed	**552,029**	**24.35**	**90.97**	**91.95**	**91.46**

TABLE II
AGE RECOGNITION EVALUATION RESULTS ON THE FG-NET DATASET.

Networks	**Params (M)**	**Mean Absolute Error ↓**
ADPF [18]	14.00	2.86
MSFCL [19]	15.00	2.71
AggerCPU [4]	0.49	2.71
BridgeNet [20]	120.00	**2.56**
MWR based on VGG16 [21]	40.00	**2.23**
Proposed	**0.56**	**2.69**

TABLE III
GENDER RECOGNITION EVALUATION RESULTS ON LFW DATASET.

Networks	**Number of Parameters**	**Accuracy (%) ↑**
Althnian et al. [22]	15,473,190	72.50
Rouhsedaghat et al. [23]	16,900	94.63
GenderBTE [7]	555,770	96.50
Greco et al. [24]	3,538,984	**98.73**
Proposed	**551,642**	**96.58**

epochs, utilizing the Adam optimizer. To dynamically adjust the learning rate based on changes in average accuracy, we implement a learning rate reduction mechanism by decreasing the rate by a factor of 0.75 after 20 epochs of stagnant average accuracy, contributing to adaptive learning throughout the training process. The hyperparameter reduction channel r on the BTE is empirically set to 4. The training utilizes an Nvidia GeForce GTX 1080Ti GPU with 11GB of memory through the Tensorflow and Keras framework. Additionally, we evaluate the FPS for both the proposed architecture and the recognizer using an Intel Core i7-9750H CPU running at 2.6 GHz with 20GB of RAM.

IV. EXPERIMENTAL RESULTS

A. Evaluation on Datasets

1) UTKFace for Simultaneous Age Group and Gender Recognition: Widely recognized in the field of facial age group and gender recognition research, UTKFace [11] comprises 23,708 facial images, each meticulously annotated with age, gender, and ethnicity details. Notably, it spans a wide age range from 0 to 116 years and features diverse images depicting variations in pose, illumination, expression, and other relevant factors. The dataset is partitioned into 80% for training and 20% for testing. As a result, shown in Table I, the proposed network, equipped with only 552,029 parameters and 24.35 MFLOPs, achieves good accuracies of 90.97%, 91.95%, and 91.46% for age, gender, and average recognition, respectively. This performance surpasses the state-of-the-art networks in simultaneous (multi-task) facial age group and gender recognition. This work also compares our proposed network with InceptionNeXt_N16 [14], VAN_N16 [17], GhostNet_N10 [15], and FasterNet_N16 [16] networks, which means applied 16, 16, 10, and 16 as initial embedding channel dimensions, respectively, to generate the networks variant that has a comparable number of parameters with our proposed model.

2) FG-NET for Age Recognition: FG-NET [25] dataset comprises 1,002 facial images obtained from 82 subjects, showcasing various variations in pose, expression, and illumination. Adhering to established protocols [26], [27], the dataset adopts k-fold cross-validation and leave-one-person-out (LOPO) methodologies. The evaluation process of this dataset computes results based on average values, utilizing the mean absolute error (MAE) metric. Table II shows the MAE result of the proposed network compared to the state-of-the-art networks. Despite securing the third position, the model demonstrates strong competitiveness, boasting an MAE of 2.69 for age recognition. This trails behind the top-ranking model by a mere 0.46 and the second-best by just 0.13. Remarkably, the proposed architecture achieves this level of performance while maintaining a significantly lower parameter count compared to its competitors.

3) LFW for Gender Recognition: The LFW [28] dataset comprises over 13,000 face images, characterized by two labels: females and males, with females accounting for 23% and males for 77% of the total instances, demonstrating a considerable imbalance. Following the previous setting [7], this dataset is divided into training (70%) and testing (30%) sets. Table III illustrates the competitive accuracy achieved by the proposed network, which reaches 96.58% as the second-best among state-of-the-art networks. However, the proposed network boasts significantly fewer parameters compared to the first-best.

B. Model Analysis

This section explores the impact of individual components within the proposed module on recognition performance using the UTKFace dataset. Initially, we conduct an ablation study

TABLE IV
ABLATION STUDY OF THE PROPOSED NETWORK FOR SIMULTANEOUS AGE GROUP AND GENDER RECOGNITION ON UTKFACE DATASET.

Baseline	**Local with Attention Modules**	**Global**	**Params**	**MFLOPs**	**Age Accuracy (%) ↑**	**Gender Accuracy (%) ↑**	**Average Accuracy (%) ↑**
✓			549,093	23.92	88.47	91.93	90.20
✓	✓		549,765	24.03	90.46	91.73	91.10
✓		✓	551,485	24.27	90.43	**92.08**	91.26
✓	✓	✓	552,029	24.35	**90.97**	91.95	**91.46**

TABLE V
COMPARISONS OF DIFFERENT BLOCKS APPLIED ON THE FOURTH STAGE OF THE PROPOSED NETWORK FOR SIMULTANEOUS AGE GROUP AND GENDER RECOGNITION ON UTKFACE DATASET.

Fourth Stage Block	**Params**	**MFLOPs**	**Age Accuracy (%) ↑**	**Gender Accuracy (%) ↑**	**Average Accuracy (%) ↑**
FasterNet Block [16]	568,293	26.34	88.92	91.93	90.43
VAN Block [17]	601,061	30.65	90.08	91.77	90.93
GhostNet Block [15]	638,917	27.03	89.92	92.08	91.00
RM2L + DELA Block from AggerCPU [4]	572,647	26.41	89.72	**92.31**	91.02
InceptionNeXt Block [14]	582,709	28.34	89.92	**92.31**	91.12
BTE Block [7]	555,645	24.80	90.26	92.11	91.19
Proposed ELGA Block	**552,029**	**24.35**	**90.97**	**91.95**	**91.46**

TABLE VI
RUNTIME EFFICIENCY OF THE SIMULTANEOUS AGE GROUP AND GENDER RECOGNITION ON UTKFACE DATASET.

Networks	**Params**	**MFLOPs**	**Age Accuracy (%) ↑**	**Gender Accuracy (%) ↑**	**Average Accuracy (%) ↑**	**AG & GD (FPS) ↑**	**FD + AG & GD (FPS) ↑**
VAN_N16 [17]	617,381	76.69	88.74	90.39	89.57	42.75	37.05
InceptionNeXt_N16 [14]	695,845	82.43	86.38	88.54	87.46	54.03	44.45
GhostNet_N10 [15]	665,765	15.43	86.62	89.16	87.89	89.25	65.61
FasterNet_N16 [16]	623,621	35.61	89.19	89.70	89.45	136.79	89.73
AggerCPU [4]	486,951	40.58	89.77	91.79	90.78	183.79	109.18
GenderBTE [7]	555,645	24.80	90.26	92.11	91.19	**199.39**	**113.46**
Proposed	**552,029**	**24.35**	**90.97**	**91.95**	**91.46**	**194.29**	**113.43**

by systematically removing each module from the proposed model and assessing the resulting performance differences. Additionally, we analyze the final feature extraction process through a late block comparison analysis.

1) Ablation Study: Table IV presents the findings from the ablation study, focusing on the average accuracy metrics for simultaneous age group and gender. The results indicate that employing the Local (with Attention Modules) and Global branches individually can improve performance by 0.90% and 1.06%, respectively, compared to the baseline consisting of the initial three stages and downsampling blocks. Furthermore, integrating Local and Global modules leads to a performance enhancement of 1.26% compared to the baseline.

2) Late Block Comparison Analysis: This analysis involves replacing the proposed ELGA block in the fourth stage with state-of-the-art alternatives. Table V shows that the proposed ELGA block consisting of the initial three stages and downsampling blocks can provide the best average accuracy of age group and gender recognition. ELGA block can outperform BTE, InceptionNeXt, and the other state-of-the-art blocks.

C. Runtime Efficiency

The offered simultaneous facial age group and gender recognition using the proposed ELGA-Net with 552,029 parameters and 24.35 MFLOPs demonstrates real-time efficiency on a CPU with an Intel Core i7-9750H 2.6 GHz, as shown in Table VI. It achieves second-best, reaching a speed of 194.29 frames per second for simultaneous facial age group and gender recognition (AG & GD) and 113.43 frames per second when integrated with face detection [29] (FD + AG & GD). This speed is slightly slower (0.03 FPS difference) than the fastest network, GenderBTE [7]. Even so, the proposed network offers superior performance based on age group and gender average accuracy (0.27% difference). Fig. 3 shows the recognition results of the proposed network, where light blue and yellow bounding boxes indicate the male and female faces, respectively.

V. CONCLUSION

This work introduces an efficient local-global attention network (ELGA-Net) for simultaneous facial age group and gender recognition. It comprises an efficient local-global attention (ELGA) block utilizing multi-scale and multi-band-level-based depthwise convolution to capture local information and a bottleneck transformer encoder to grasp global context and relationships between different parts of the feature maps in a parallel structure. The proposed network demonstrates competitive performance on the UTKFace dataset for simultaneous facial age group and gender recognition, and it also performs well on the FG-NET and LFW datasets for separate age group and gender recognition tasks. Additionally, the proposed network achieves real-time speed at 113.43 FPS on a CPU device, integrated with face detection as an initial process.

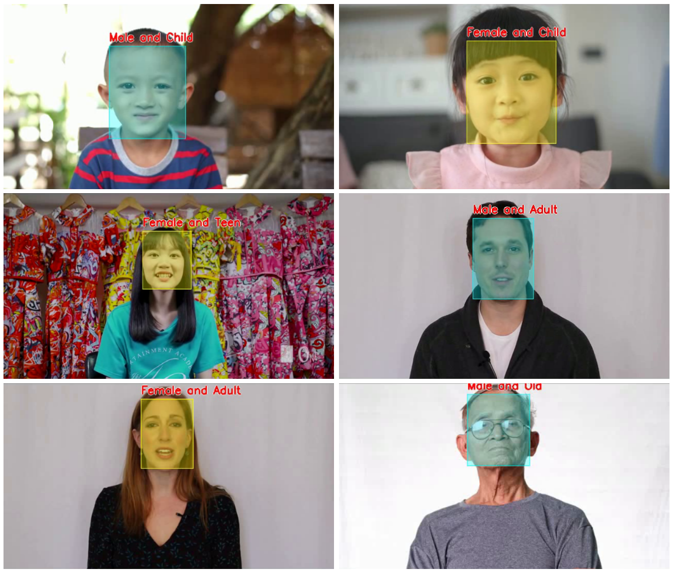

Fig. 3. The simultaneous facial age group and gender recognition results on new samples of the proposed network trained on the UTKFace dataset.

In future work, the proposed recognition will be extended to operate on more cost-effective devices to support Robot Vision applications.

Acknowledgment

This result is 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] N. Kuratomo, H. Miyakawa, T. Ebihara, N. Wakatsuki, K. Mizutani, and K. Zempo, "Attracting effect of pinpoint auditory glimpse on digital signage," *IEEE Access*, 2023.

[2] L. Wang, Z. Yu, D. Yang, H. Ma, and H. Sheng, "Efficiently targeted billboard advertising using crowdsensing vehicle trajectory data," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1058–1066, 2019.

[3] A. Priadana, M. D. Putro, J. An, D.-L. Nguyen, X.-T. Vo, and K.-H. Jo, "Facial attribute recognition using lightweight multi-label cnn-transformer architecture for intelligent advertising," in *IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2023, pp. 1–7.

[4] A. Priadana, M. D. Putro, D.-L. Nguyen, X.-T. Vo, and K.-H. Jo, "Age group recognizer based on human face supporting smart digital advertising platforms," in *2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)*. IEEE, 2023, pp. 1–7.

[5] C. Wang, Z. Li, X. Mo, X. Tang, and H. Liu, "Exploiting unfairness with meta-set learning for chronological age estimation," *IEEE Transactions on Information Forensics and Security*, 2023.

[6] M. Shahzeb, S. Dhavale, D. Srikanth, and S. Kumar, "Dcnn-based transfer learning approaches for gender recognition," in *International Conference on Data Management, Analytics & Innovation*. Springer, 2023, pp. 357–365.

[7] A. Priadana, M. D. Putro, J. An, D.-L. Nguyen, X.-T. Vo, and K.-H. Jo, "Gender recognizer based on human face using cnn and bottleneck transformer encoder," in *2023 International Workshop on Intelligent Systems (IWIS)*. IEEE, 2023, pp. 1–6.

[8] M. D. Putro, A. Priadana, D.-L. Nguyen, and K.-H. Jo, "A faster real-time face detector support smart digital advertising on low-cost computing device," in *2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 2022, pp. 171–178.

[9] H. Liao, L. Yuan, M. Wu, L. Zhong, G. Jin, and N. Xiong, "Face gender and age classification based on multi-task, multi-instance and multi-scale learning," *Applied Sciences*, vol. 12, no. 23, p. 12432, 2022.

[10] P. Vidyarthi, S. Dhavale, and S. Kumar, "Gender and age estimation using transfer learning with multi-tasking approach," in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*. IEEE, 2022, pp. 1–5.

[11] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4352–4360.

[12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 11 966–11 976.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] W. Yu, P. Zhou, S. Yan, and X. Wang, "Inceptionnext: when inception meets convnext," *arXiv preprint arXiv:2303.16900*, 2023.

[15] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 1577–1586.

[16] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 021–12 031.

[17] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Computational Visual Media*, vol. 9, no. 4, pp. 733–752, 2023.

[18] H. Wang, V. Sanchez, and C.-T. Li, "Improving face-based age estimation with attention-based dynamic patch fusion," *IEEE Transactions on Image Processing*, vol. 31, pp. 1084–1096, 2022.

[19] M. Xia, X. Zhang, L. Weng, Y. Xu *et al.*, "Multi-stage feature constraints learning for age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2417–2428, 2020.

[20] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "Bridgenet: A continuity-aware probabilistic network for age estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1145–1154.

[21] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Moving window regression: A novel approach to ordinal regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 760–18 769.

[22] A. Althnian, N. Aloboud, N. Alkharashi, F. Alduwaish, M. Alrshoud, and H. Kurdi, "Face gender recognition in the wild: an extensive performance comparison of deep-learned, hand-crafted, and fused features with deep and traditional models," *Applied Sciences*, vol. 11, no. 1, p. 89, 2020.

[23] M. Rouhsedaghat, Y. Wang, X. Ge, S. Hu, S. You, and C.-C. J. Kuo, "Facehop: A light-weight low-resolution face gender classification method," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 169–183.

[24] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff," *IEEE Access*, vol. 8, pp. 130 771–130 781, 2020.

[25] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[26] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, "Deep differentiable random forests for age estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 404–419, 2019.

[27] H. Liu, J. Lu, J. Feng, and J. Zhou, "Label-sensitive deep metric learning for facial age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 2, pp. 292–305, 2017.

[28] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[29] M. D. Putro, D.-L. Nguyen, and K.-H. Jo, "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)*. IEEE, 2020, pp. 94–99.