An Efficient Face-based Age Group Detector on a CPU using Two Perspective Convolution with Attention Modules

Adri Priadana, Muhamad Dwisnanto Putro, Xuan-Thuy Vo and Kang-Hyun Jo
Department of Electrical, Electronic, and Computer Engineering
University of Ulsan
Ulsan, Korea
priadana3202@mail.ulsan.ac.kr, dputro@mail.ulsan.ac.kr, xthuy@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

***Abstract*—Age detection has become incredibly substantial in various scenarios such as video surveillance, forensic applications, and advertising platform. An age detector is expected to operate on low-cost devices or CPU devices to minimize the budget of the implementation system. This work presents an efficient face-based age group detector (Age-CPU) that can operate fluidly on a CPU. It proposes two perspectives convolution architecture with depthwise global attention modules (2PDG) on this detector. It applies two kernel sizes to consider different sizes of the feature area of the object reinforced with enhancing block. The depthwise layer on the attention module helps the architecture extract features more focused and deeply. It convolves each channel with an individual depthwise kernel. The architecture is trained and validated on the UTKFace and FG-NET datasets. 2PDG acquires competitive accuracy compared to other competitors' architectures on the datasets. Furthermore, the proposed detector can operate 100 frames per second on a CPU device, which is speedy to execute in real-time.**

***Index Terms*—face-based age group, efficient detector, attention module, real-time detector, a CPU device**

I. Introduction

Age detection technology has attracted a significant number of researchers. The objective of this technology is to estimate age group [1]–[3] or even exact age [4] as one of the facial attributes of a subject. The age group is obtained by dividing the full age range into several aging groups [2]. It is used as a target class in predicting the desired age group for various purposes, such as differentiating between adults and children for restriction purposes. Face-based age detection can be conducted by analyzing a face detected by a camera. Age detection has become incredibly substantial in various scenarios such as video surveillance, forensic applications [5], and advertising platforms [6]. Face-based age detectors commonly consist of a face detector and an age estimator. The face detector detects the face and obtains the face area from the images. The age estimator estimates the age based on the detected face area. An age estimator consists of a feature extractor and a classifier in modern techniques. Firstly, it is used to extract age features from the face. Secondly, it predicts the age or age group based on the extracted features.

The Convolutional Neural Network (CNN) is a technique used in many works [4], [7], [8] to implement age estimation and has been proven to show good performance. Chen et al. [9] used a CNN architecture as a baseline for age estimation work. They modified the original AlexNet model and proposed a novel architecture called Attribute-Region Association Network (ARAN), generating 414 million parameters. Li et al. [8] proposed a CNN architecture called BridgeNet for age estimation consisting of local regressors and gating networks. The local regressors split the data space to settle heterogeneous data, and gating networks learn aware continuity weights. The BridgeNet architecture generates 120 million parameters.

As time goes by, many works aim to design a more efficient CNN architecture with fewer parameters to create a more efficient CNN architecture. Shen et al. [10] used the CNN technique integrated with deep differentiable random forests methods to perform age estimation. It generates only 14 million parameters. Another work [4] proposed an attention-based CNN architecture called attention-based dynamic patch fusion (ADPF). It consists of two separate CNN, the AttentionNet and the FusionNet model. The first model is used to dynamically locate and rank age-specific patches, and the second uses the discovered patches to predict the subject's age. The ADPF also generates only 14 million parameters. These two works show that the CNN architecture they developed has fewer the number of a parameter, which makes it more efficient.

The age detector is expected to run on low-cost devices or CPU devices to reduce the costs of the implementation system. Therefore, the light CNN architecture with very few parameters is required, making the detector more efficient and compatible with running in real-time on low-cost or CPU devices. The CNN architecture with a few parameters also makes the detector run faster. This work presents an efficient real-time face-based age group detector with a low parameter appropriate for a CPU device.

An efficient face-based age group detector (Age-CPU) proposed two perspectives convolution architecture with depthwise global attention modules (2PDG) that offer two view contextual ways reinforced with attention modules. It applies two kernel sizes to reckon different sizes of the feature area of the object. The architecture only produces a few parameters that make the detector efficient and run faster. Therefore, the

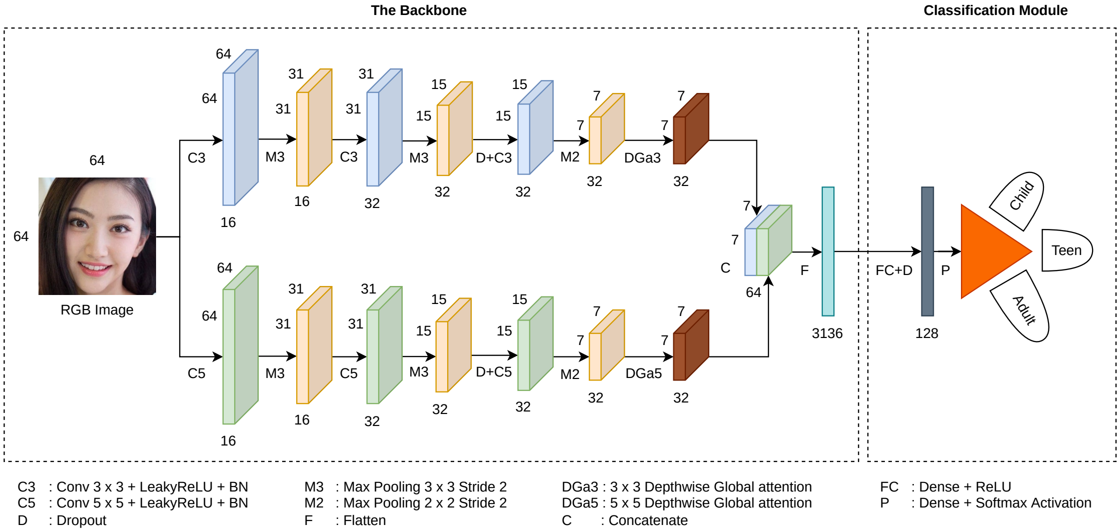

The Backbone
Classification Module
64
64
RGB Image
C3
C5
M3
D+C3
D+C5
M2
DGa3
DGa5
C
F
FC+D
P
3136
128
Child
Teen
Adult
C3 : Conv 3 x 3 + LeakyReLU + BN
C5 : Conv 5 x 5 + LeakyReLU + BN
D : Dropout
M3 : Max Pooling 3 x 3 Stride 2
M2 : Max Pooling 2 x 2 Stride 2
F : Flatten
DGa3 : 3 x 3 Depthwise Global attention
DGa5 : 5 x 5 Depthwise Global attention
C : Concatenate
FC : Dense + ReLU
P : Dense + Softmax Activation

Fig. 1. The proposed architecture of the efficient face-based age group detector. It uses two perspectives convolution architecture with depthwise global attention modules.

face-based age detector can be convenient for low-cost or CPU devices. The contribution of this work summarizes as follows:

1) An efficient CNN backbone using two perspectives convolution architecture with squeezed kernels is proposed to extract features with two diverse kernel sizes. Both run detachedly during the feature extraction component reinforced with attention modules. It can capture adequate quality information from distinct feature areas of the object.
2) A new depthwise global attention module is proposed as an escalation module to improve the quality of the feature map from the input feature. The performance result gains competitive accuracy with other architectures on UTKFace [11] and FG-NET [12] datasets.
3) An efficient face-based age group detector is offered that is suitable for implementation on a CPU device. The performance result gains competitive speed with other common and light CNN architectures.

II. Proposed Architecture

The proposed architecture employs two convolution layer sequences that run parallel with depthwise global attention modules, as shown in Fig. 1. In this architecture, the backbone efficiently extracts features of faces, and the classification module predicts the age group of the face. The proposed architecture of this work generates 459,347 parameters.

A. The Backbone

The Age-CPU proposes a backbone using two perspectives convolution architecture with depthwise global attention modules to extract age features from the face image. Inspired by the backbone in [13], this backbone of the proposed architecture only consists of two sequences of convolution layers, the primary key for extracting age features from a face. It applies different kernel sizes to capture features based on local dependencies in various areas in each sequence. We name them as different perspectives in viewing. It aims to obtain more information and enrich the spatial component extracted from the images. Unlike in [13], it only uses 3 × 3 and 5 × 5 kernel sizes in this backbone. Each perspective consists of three convolution layers arranged sequentially with one-times expansion in the number of kernels from 16, 32, and 32. Leaky ReLU (Leaky Rectified Linear Unit) activation function and a batch normalization technique [14] are used after convolution operations to bargain with the gradient problem. It also applies a dropout technique before the last convolution layer to prevent overfitting [15].

In order to shrink the feature map and summarize the essential features with high activation values, it applies max-pooling operations with different sizes. After the first and second convolution layers, it put a 3 × 3 max-pooling layer with strides two to summarize the broader area in this architecture's low and middle-level features. Further, it put a 2 × 2 max-pooling layer with stride two after the last convolution layer. The attention module is assigned to improve the quality of the feature map produced by the previous layers, as will be discussed in detail in the next section. Then, a concatenate operation is applied to fuse the two perspectives, followed by a flattening process to make a one-dimensional vector and feed them to the classification module.

B. The Depthwise Global Attention Module (DG)

In the CNN technique, an attention mechanism is an approach of selectively focusing on a few features of the images while ignoring others. Expanding the global attention mechanism in [16], we not only perform a global average-pooling operation to aggregate each feature map but also perform a global max-pooling operation in a parallel manner as shown in Fig. 2. It aims to strengthen the exactitude in selecting interest features because it globally highlights each

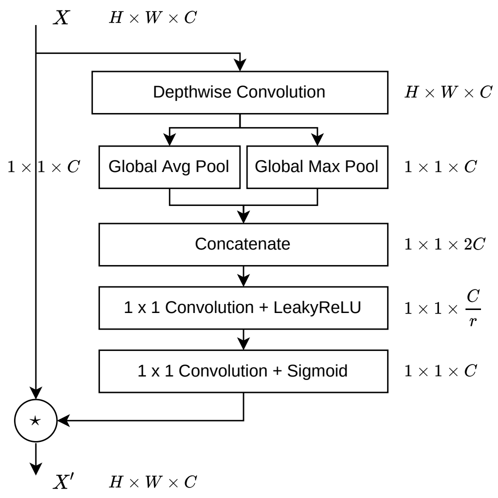


Fig. 2. The proposed depthwise global attention module. It is used to improve the quality of the feature map.

channel and uses it as a basis to determine which channels have rich interest features. It produces two vectors representing the feature summary of the corresponding channel. Further, a concatenation operation is applied to fuse these vectors.

In order to provide an opportunity on each channel to deepen learning without being impacted by information from other channels before summarizing, depthwise convolution is applied before the global pooling operations. It only convolves each channel with an individual depthwise kernel. After the concatenation operation, two sequential 1 × 1 convolution operations are applied to capture channel-wise dependencies fully. It will also capture the dependencies of two aggregation operations results. A dimensionality-reduction mechanism is used with a reduction ratio r at the first convolution layer to make it more efficient. A Leaky ReLU activation is also used after the first convolution layer to avoid the loss of valuable information because the activation considers the positive and negative values. The proposed depthwise global attention module equation is expressed as:

$$d(x, k) = W_D(x, k), \tag{1}$$

$$A(x, d) = x * \sigma(W_{C2}(\delta(W_{C1}(Co\,[Ga(d), Gm(d)])))), \tag{2}$$

where x is the input of the attention module, d is the output of the depthwise convolution layer, W_D is learnable parameters in the depthwise convolution layer, k is the kernel size applied in the depthwise convolution layer, Ga is the global average-pooling operation, Gm is the global max-pooling operation, Co is the concatenation operation, δ refers to the Leaky ReLU activation function, W_{C1} and W_{C2} are learnable parameters in the two 1 × 1 convolution layers, and σ indicates the Sigmoid function used to normalize the attention weights. The kernel size of the depthwise convolution layer depends on which perspective the attention module will be applied. The kernel size is 3 × 3 or 5 × 5.

C. Classification Module

This module is used to compute the probability of the age group class to predict the age group of the face detected by the camera. It consists of two dense layers. The first layer consists of 128 units with ReLU (Rectified Linear Unit) activation. The second layer consists of the Softmax activation function that generates the input vector to possibilities representing the prediction result class. The equation of the Softmax activation function is expressed as:

$$S\left(z_i\right) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \left(i = 1, 2, ..., N\right), \tag{3}$$

where $z_1, z_2, ..., z_N$ are the input values of the Softmax layer and the output value $S\left(z_i\right)$ represents the probability that the sample belongs to the i-th class. In order to prevent overfitting, a dropout operation is also performed before the second layer.

D. Face Detector

The face-based age group detector requires face detection as a preliminary process to obtain the face area as a Region of Interest (RoI) and feed it as an input for the detector architecture. In order to support the performance of the face-based age group detector, a face detector with efficient performance is required, mainly to perform in real-time scenarios. Therefore, the LWFCPU [17] face detector with light architecture is utilized in this work. It only applies twelve convolutional layers and generates a few parameters capable of running fast in real-time on low computing or CPU devices.

III. Implementation Setup

The proposed architecture is implemented on Keras 2.3.1 and the Tensorflow 2.0 framework. It trained on the NVIDIA Tesla V100-PCIe 32GB as an accelerator and tested on Intel Core i7-9750H CPU @ 2.60GHz with 20GB RAM. The training and validation process is conducted on the UTKFace and FG-NET datasets with 300 number of epoch in the training stage. The initial learning rate is set to 10^{-3} and will be reduced to 75% when the accuracy does not improve every 20 epochs. The training uses a batch size of 256. Moreover, Adam is used as an optimizer to update the weight based on Categorical Cross-Entropy loss. This implementation setup is applied to all the dataset settings of this work.

IV. Experimental Results

The examination result of the proposed architecture on the datasets benchmark will be described in this section. The runtime efficiency, limitation, ablative study, and attention modules comparison are also described in this section.

A. Evaluation on Datasets

1) UTKFace (Aligned and Cropped Faces): The dataset contains more than 23,000 facial images covering many variations such as expression, illumination, age, resolution, pose, etc. The age variations range from 0 to 116. Three commonly used settings and one proposed setting are adopted for this dataset evaluation. In the first setting, i.e., *Setting I*, following prior work [18], the dataset is divided into the two subsets with a random permutation split, 90% as training and 10% as testing sets. It will randomly reorder a collection of objects in a different order than the original or previous order. In this setting, five age groups are used as age class targets, 0-24, 25-49, 50-74, 75-99, and 100-116. In the second setting, i.e., *Setting II*, following prior work [19], the dataset is divided into the three subsets with a random permutation split, 80% as training, 10% as validation, and 10% as testing sets. In this setting, the same five groups as Setting I are used.

In the third setting, i.e., *Setting III*, following prior work [20], the dataset is divided into the three subsets with a random permutation split, 10,437 images as training, 3,252 images as a validation, and 10,719 images as testing. In this setting, seven age groups are used as age class targets, 0-3 as baby's face, 4-12 as child's face, 13-19 as teenager's face, 20-30 as young's face, 31-45 as adult's face, 46-60 as middle-aged's face, and 61-116 as senior's face. The fourth setting, i.e., *Setting IV*, is used for our detector. This detector aims to distinguish between a child's and an adult's faces. In order to avoid an extreme separation between two classes, the dataset is divided into three age groups as age class targets, 0-11 as child's face, 12-17 as teen's face, and 18-116 as adult's face. It is also divided into the two subsets with a random permutation split, 90% as training and 10% as testing sets.

In this dataset, results are conveyed based on two metrics. They are Mean Absolute Error (MAE) for the Setting I and Validation Accuracy (VA) for the Setting II-IV, which are shown in Table I and Table II, respectively. As can be seen in Setting I, 2PDG with only 459,605 parameters outperforms all competitor architectures that applied transfer learning mechanism. In Setting II-III, 2PDG achieves competitive performance based on the validation accuracy of the other architectures. For Setting II, 2PDG result is below [19], which differs only by 0.43, but it only generated about 50% less number of parameters. In Setting IV, the proposed architecture achieves the best performance concerning validation accuracy of the common architectures, especially of the two light architectures such as MobileNetV2 and SqueezeNet, which differed by 0.33 and 1.81, respectively. The proposed architecture also produces more efficiency than the other architectures according to the number of parameters.

2) FG-NET: The dataset comprises 1,002 facial images from 82 subjects covering illumination, expression, and pose variations. Every subject of this dataset has more than ten facial images. Following prior work [10], [21], it uses the leave-one-person-out (LOPO) and k-fold cross-validation techniques in this dataset. In every fold, facial images of one subject are used for testing and the rest for the training process. There are 82 subjects on this dataset.Therefore, this evaluation process implements 82 fold, and the conveyed results are the average values. The evaluation result of this dataset reports according to the MAE metric, shown in Table III. It can be seen that 2PDG with the training setting achieves competitive performance and occupies the third-best ranking with 2,75 MAE, which differs only by 0.19 and 0.02 from the best and the second-best, respectively. Even so, the proposed architecture generates total parameters far below the competitors. Moreover, 2PDG reaches an MAE value under 3.00. It indicates that it can execute adequately even with a small dataset.

TABLE I
EVALUATION RESULTS ON UTKFACE SETTING I (FIVE AGE GROUPS) DATASET

Architectures	**Number of Parameters**	**MAE**
ResNet50 with Transfer Learning [18]	23,597,957	9.66
InceptionV3 with Transfer Learning [18]	21,813,029	9.50
DenseNet with Transfer Learning [18]	7,042,629	9.19
2PDG	**459,605**	**8.55**

TABLE II
EVALUATION RESULTS ON UTKFACE SETTING II-IV DATASET

Architectures	**Number of Parameters**	**VA (%)**
UTKFace Setting II (Five Age Groups)		
Best CNN [19]	963,069	**79.12**
2PDG	**459,605**	**78.69**
UTKFace Setting III (Seven Age Groups)		
Facenet [20]	-	56.90
FFNet [20]	-	64.00
MTCNN [20]	-	**70.10**
2PDG	**459,863**	**65.07**
UTKFace Setting IV (Three Age Groups)		
SqueezeNet + Batch Normalization	735,823	94.60
InceptionV3	21,808,931	95.06
ResNet50V2	23,570,947	95.10
VGG16 + Batch Normalization	39,786,819	95.27
VGG11 + Batch Normalization	34,417,795	95.48
VGG13 + Batch Normalization	34,472,003	95.82
MobileNetV2	2,261,827	96.08
2PDG	**459,347**	**96.41**

B. Runtime Efficiency and Limitation

Age-CPU with 2PDG architecture is designed to operate on low-cost devices or CPU devices to minimize the budget of the implementation system. With only 459,347 parameters, 2PDG can perform efficiently in real-time on CPU-based. It achieves 166 FPS in classifying face-based age group (Age Group) and 100 FPS in recognizing facial age group when integrated with face detection (Face + Age Group). 2PDG becomes the most rapid detector on the CPU compared to other competitors, as shown in Table IV. Fig. 2 (a) shows the correct prediction results of the Age-CPU detector on the CPU. The green bounding box indicates a child's face, the blue bounding box indicates a teen's face, and the red indicates an adult's face. As shown in Fig. 2 (b), the proposed detector is still weak in predicting the age group on the face with yaw pose because the UTKFace dataset does not have many instances, especially on the face with yaw pose.

C. Ablative Study and Attention Modules Comparison

In this experiment, the performance of each proposed module is investigated by removing each of them, then measuring

TABLE III
EVALUATION RESULTS ON FG-NET DATASET

Architectures	Loss Function	Optimizer	Number of Parameters	MAE
DEX [22]	Euclidean	-	120 M	4.63
Mean-Variance Loss [7]	Mean-Variance	SGD	20 M	4.10
GA-DFL [23]	-	-	138 M	3.93
LSDML [21]	-	-	44 M	3.92
ARAN [9]	Manually-designed	-	414 M	3.79
M-LSDML [21]	-	-	44 M	3.74
DLDLF [10]	Manually-designed	-	14 M	3.71
DRF [10]	Manually-designed	-	14 M	3.41
DAG-VGG16 [24]	-	-	24 M	3.08
DAG-GoogleNet [24]	-	-	131 M	3.05
ADPF [4]	Diversity & Age Estimation	SGD	14 M	2.86
BridgeNet [8]	Regression & KL Divergence	SGD	120 M	**2.56**
VGG16 + Batch Normalization	Mean Absolute Error	SGD	40 M	3.11
VGG16 + Batch Normalization	Categorical Cross Entropy	SGD	40 M	3.03
VGG16 + Batch Normalization	Mean Absolute Error	Adam	40 M	3.01
VGG16 + Batch Normalization	Categorical Cross Entropy	Adam	40 M	**2.73**
MobileNetV2	Categorical Cross Entropy	Adam	2.34 M	3.12
SqueezeNet + Batch Normalization	Categorical Cross Entropy	Adam	0.77 M	3.05
2PDG	**Categorical Cross Entropy**	**Adam**	**0.46 M**	**2.75**

TABLE IV
COMPARISON OF ARCHITECTURE SPEEDS ON A CPU

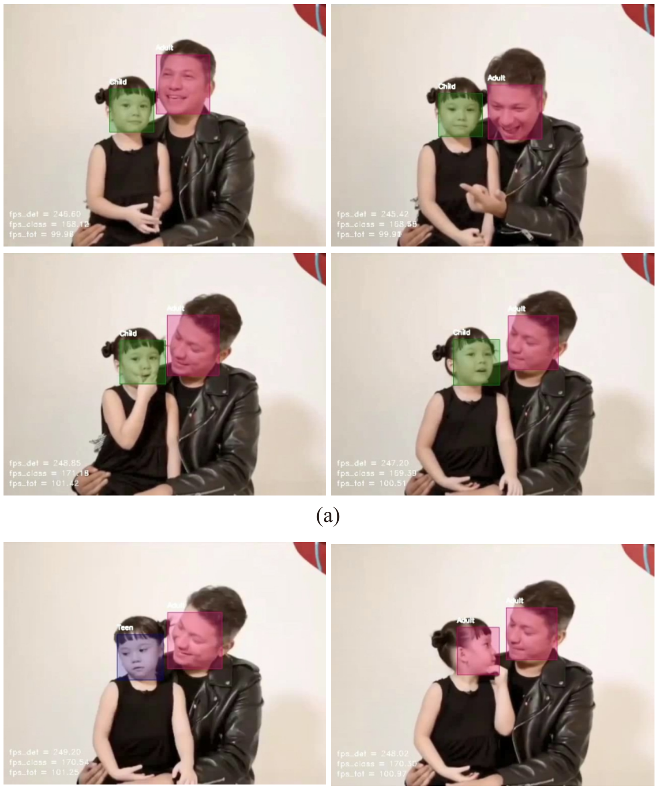
Architectures	Age Group (FPS)	Face + Age Group (FPS)
InceptionV3	31	27
ResNet50V2	35	32
VGG16 + Batch Normalization	38	33
VGG13 + Batch Normalization	45	38
VGG11 + Batch Normalization	50	41
MobileNet V2	56	46
Squeezenet + Batch Normalization	99	71
2PDG	**166**	**100**

TABLE V
ABLATIVE STUDY OF PROPOSED ARCHITECTURE

Settings	Number of Parameters	VA (%)	Age Group (FPS)	Face + Age Group (FPS)
2P	456,579	95.49	201	113
2PG	458,195	95.95	170	103
2PDG	459,347	96.41	166	100

its performance and efficiency. It will reveal the influence of the presence of each proposed module. It uses the UTKFace Setting IV dataset in this investigation. According to the VA metric, the result reports are shown in Table V. It can be seen that utilizing the proposed global attention module (2PG) can increase classification capability by 0.46%. Further, adding the depthwise convolution layer at the global attention module (2PDG) can increase classification capability by 0.46%. In addition, the use of the modules does not significantly reduce the efficiency of the architecture, which only decreases 13 FPS.

The proposed attention module with various reduction ratios r is also compared with other lightweight attention methods such as Squeeze-and-Excitation (SE) [25] and Convolutional Block Attention Module (CBAM) [26], which is illustrated in Table VI. In order to perform a fair comparison, it applies the attention module in the proposed backbone (2P). The validation accuracy of the proposed attention module, DG (r=4) is higher than SE (r=4) and CBAM (r=4), which differ



(a)

(b)

Fig. 3. The correct prediction result (a) and the incorrect prediction results (b) of the Age-CPU detector.

TABLE VI

COMPARISONS OF DIFFERENT ATTENTION METHODS WHEN USING PROPOSED TWO PERSPECTIVE NETWORK (2P) AS THE BACKBONE

Attention Modules	Number of Parameters	VA (%)	Age Group (FPS)	Face + Age Group (FPS)
SE (r=4) [25]	457,603	95.70	181	106
DG (r=8)	458,571	95.82	169	101
DG (r=2)	460,899	96.12	164	100
CBAM (r=4) [26]	457,719	96.20	128	85
DG (r=4)	**459,347**	**96.41**	**166**	**100**

by 0.71% and 0.21%, respectively. Based on speed, 2PDG ranks second best after SE, with a difference of only 6 FPS when integrated with face detection. CBAM ranks the lowest in terms of speed because this module consists of a channel and a spatial attention module.

V. CONCLUSION

This study proposes an efficient real-time face-based age group detector with lightweight architecture. It offers two perspectives convolution architecture with depthwise global attention modules (2PDG). The proposed depthwise global attention module is used to improve the quality of the feature map resulting from the previous operation. The 2PDG gained competitive accuracy compared to other competitors on the UTKFace and FG-NET datasets. As a result, the detector can operate at 100 FPS to recognize the age group of the face when working on a CPU device in real-time. The proposed attention module also gains the best performance compared with other attention modules such as SE and CBAM. In future work, the dataset will be explored more thoroughly to address the limitations of the proposed detector. It is also potential for future work to combine age estimation straight into the face detection network to become one package detector.

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] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li, “Age group and gender estimation in the wild with deep ror architecture,” *IEEE Access*, vol. 5, pp. 22 492–22 503, 2017.

[2] M. T. B. Iqbal, M. Shoyaib, B. Ryu, M. Abdullah-Al-Wadud, and O. Chae, “Directional age-primitive pattern (dapp) for human age group recognition and age estimation,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2505–2517, 2017.

[3] A.-T. Mai, D.-H. Nguyen, and T.-T. Dang, “Real-time age-group and accurate age prediction with bagging and transfer learning,” in *2021 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, 2021, pp. 27–32.

[4] H. Wang, V. Sanchez, and C.-T. Li, “Improving face-based age estimation with attention-based dynamic patch fusion,” *IEEE Transactions on Image Processing*, 2022.

[5] F. Becerra-Riera, A. Morales-González, and H. Méndez-Vázquez, “A survey on facial soft biometrics for video surveillance and forensic applications,” *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1155–1187, 2019.

[6] S. Suman and S. Urolagin, “Age gender and sentiment analysis to select relevant advertisements for a user using cnn,” in *Data Intelligence and Cognitive Informatics*. Springer, 2022, pp. 543–557.

[7] H. Pan, H. Han, S. Shan, and X. Chen, “Mean-variance loss for deep age estimation from a face,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5285–5294.

[8] 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.

[9] Y. Chen, S. He, Z. Tan, C. Han, G. Han, and J. Qin, “Age estimation via attribute-region association,” *Neurocomputing*, vol. 367, pp. 346–356, 2019.

[10] 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.

[11] Z. Zhang, Y. Song, and H. Qi, “Age progression/regression by conditional adversarial autoencoder,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5810–5818.

[12] 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.

[13] A. Priadana, M. D. Putro, and K.-H. Jo, “An efficient face gender detector on a cpu with multi-perspective convolution,” in *2022 13th Asian Control Conference (ASCC)*, 2022, pp. 453–458.

[14] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[17] 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.

[18] M. Akhand, I. Sayim, S. Roy, N. Siddique *et al.*, “Human age prediction from facial image using transfer learning in deep convolutional neural networks,” in *Proceedings of International Joint Conference on Computational Intelligence*. Springer, 2020, pp. 217–229.

[19] V. Sheoran, S. Joshi, and T. R. Bhayani, “Age and gender prediction using deep cnns and transfer learning,” in *International Conference on Computer Vision and Image Processing*. Springer, 2020, pp. 293–304.

[20] A. Das, A. Dantcheva, and F. Bremond, “Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach,” in *Proceedings of the european conference on computer vision (eccv) workshops*, 2018, pp. 0–0.

[21] 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.

[22] R. Rothe, R. Timofte, and L. Van Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” *International Journal of Computer Vision*, vol. 126, no. 2, pp. 144–157, 2018.

[23] H. Liu, J. Lu, J. Feng, and J. Zhou, “Group-aware deep feature learning for facial age estimation,” *Pattern Recognition*, vol. 66, pp. 82–94, 2017.

[24] S. Taheri and Ö. Toygar, “On the use of dag-cnn architecture for age estimation with multi-stage features fusion,” *Neurocomputing*, vol. 329, pp. 300–310, 2019.

[25] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2019.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.