2022 The 22st International Conference on Control, Automation and Systems (ICCAS 2022)
BEXCO, Busan, Korea, Nov. 27∼Dec. 01, 2022

A Facial Gender Detector on CPU using Multi-dilated Convolution with Attention Modules

Adri Priadana[1*], Muhamad Dwisnanto Putro[2], Xuan-Thuy Vo[3], and Kang-Hyun Jo[4]

[1,2,3,4]Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, Ulsan, Korea, (priadana3202@mail.ulsan.ac.kr) * Corresponding author

Abstract: Facial gender detectors have evolved into a vital component of an intelligent advertisement display platform. It is helpful to assist a decision of delivering appropriate advertisements to each audience. To reduce system costs, applications deployed in this platform must be able to run on a CPU. This work proposes a facial gender detector (FG-CPU) that can be implemented on a CPU device to support an advertising display platform. The proposed CNN model consists of a multi-dilated convolution with attention modules (MudaNet). The multi-dilated convolution is applied to capture multi-scale features in an efficient manner. The attention module is used to rectify the quality of the feature map. This work's training and validation process is conducted on the UTKFace, the Labeled Faces in the Wild (LFW), and the Adience Benchmark datasets. As a result, the proposed CNN model is proven to compete with other common and lightweight competitors' CNN models on these three datasets. Regarding speed, the detector can operate 49.19 frames per second in real-time on a CPU device.

Keywords: facial gender detector, smart digital advertising, convolutional neural network, dilated convolution, attention module

1. INTRODUCTION

Smart advertisement displays have been widely arising in recent years. They can be seen in public places like airports, markets, and hotels [1]. Practically, advertisement displays have the advantage of dynamic content that can be easily personalized and customized. However, the market demands an improved approach to delivering targeted ads to the audience [2]. It can be achieved by analyzing the audience's attributes facing the platform.

Gender is one of the demographic attributes that the platforms can utilize to segment the audience [3]. By recognizing the gender, the advertisement display platform can deliver more relevant ads for each audience [4, 5]. Recognizing gender is performed by detecting and classifying a face in real-time through a camera mounted on the platform. In its application, an advertisement display platform demand a low-cost device to reduce the implementation cost [6, 7]. Therefore, it requires a facial gender detector that can be properly performed on a low-cost device or a CPU.

The deep learning technique based on Convolutional Neural Network (CNN) is extending rapidly and has many successes in recognition technology. Various CNN models have been developed to build recognition systems, especially in recognizing facial gender. In [8], a CNN model, namely HyperFace-ResNet, has been proposed to predict gender based on a face. It reached 94% and 98% accuracy on LFW and CelebA datasets, respectively. It used ResNet as a baseline that applied a shortcut connections mechanism between the lower and deeper layers. Another work [9] proposed a CNN model to predict gender based on a face and reached 89.97% accuracy on the UTKFace dataset.

The utilization of CNN to recognize gender through a face has also been applied in advertising displays in previous work. In [1], the MobileNetV2 model was utilized to develop a real-time gender detector. It was implemented on a CPU device in the advertising displays platform. The model produces 3.5 million number of parameters, which is still quite a lot, although it employs a mobile version of the CNN model. Another work [10] designed an efficient CNN model called MPConvNet that proposed a multi-perspective convolution with various kernel sizes. It only produces 659,650 parameters and reached 92.32% accuracy on UTKFace. It can run 38.72 frames per second when integrated with face detection. The facial gender detector will operate more effectively and quickly with fewer parameters. This work presents a facial gender detector with a light parameter that can perform gender recognition efficiently.

A facial gender detector (FG-CPU) presents a multi-dilated convolution with attention modules (MudaNet). The multi-dilated convolution is applied to capture multi-scale features in an efficient manner. The attention module is employed to escalate the quality of the feature map resulting from the convolution operation in the previous layer. The model generates few parameters and presides the detector to run speedily. Therefore, this model can be applied to CPU-based or low-cost devices. The main contribution of this work outlines as follows:

1. A multi-dilated convolution model with attention modules (MudaNet) is promoted, which generates few parameters and makes the model efficient. The attention modules can strengthen the face's essential features that increase the accuracy of the facial gender recognition outcome.
2. A facial gender detector is introduced, which can operate speedily on a CPU device in real-time. The proposed CNN model's performance is proven to compete with other general and lightweight CNN models on UTKFace [11], Labeled Faces in the Wild (LFW) [12], and Adience Benchmark [13] datasets.

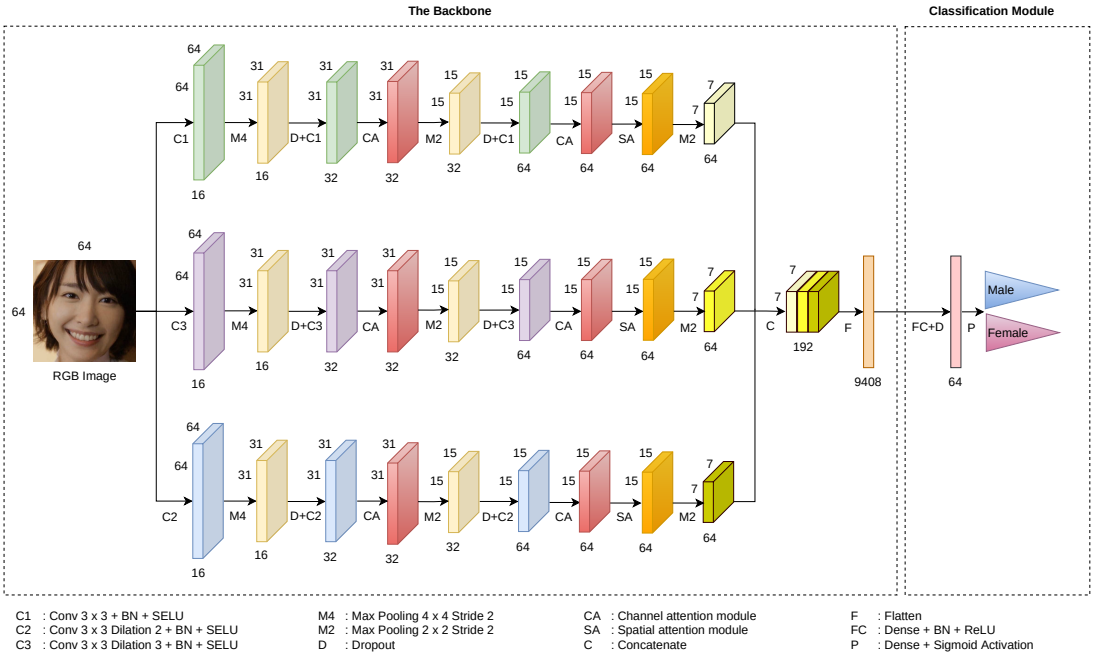

Fig. 1. The proposed CNN model of the facial gender detector. It implement a multi-dilated convolution with channel and spatial attention modules to extract features of a face.

2. PROPOSED MODEL

The proposed CNN model employs three branches of convolution layers sequence with attention modules, as seen in Fig. 1. It consists of a backbone and classification module. The backbone consists of multi-dilated convolutions and attention modules, which extract face features, and the classification module predicts the gender. It generates 674,760 parameters.

2.1 The Backbone

The FG-CPU offers a backbone module consisting of three perspectives convolution layers sequence with attention modules. It is used to extract face features. Inspired by the backbone in [10], each perspectives or branch consists of three 3×3 convolution layers. Unlike in [10], it uses different dilatation rates. The first branch uses dilation rate 1 (no dilation), the second uses dilation rate 2, and the last uses dilation rate 3. In order to fetch more information at a higher level of the feature map, the 3×3 convolution layers in every branch are arranged sequentially with a two-times kernel increasing from 16, 32, and 64. It applies a few kernel numbers to press the number of parameters. It will make the model more efficient.

In this model, a batch normalization (BN) procedure [14] and Scaled Exponential Linear Units (SELU) activation [15] are used after every convolution layer to bargain with the gradient issue. The dropout mechanism is also put in previous to the second and last convolution layers to avoid overfitting [16]. Further, it applies a max-pooling operation with two strides to shrink the feature map. We put a max-pooling layer with 4×4 sizes after the first convolution layer and 2×2 sizes after the second and the third convolution layer. The 4×4 sizes objectives to recap the wider area in the low-level feature of the proposed model.

2.2 The Attention Modules

An attention technique is a strategy implemented in deep learning for selectively focusing on specific features of the images while disregarding others. Inspired by the attention technique in [17], we propose channel and spatial attention modules that only use the global average-pooling function to make the operation more efficient. In the proposed channel attention module, the global average-pooling operation aims to aggregate spatial information of each feature map. It generates a feature vector that represents the feature outline of the related channel. Further, layer normalization (LN) [18] is used to stabilize the distributions of layer inputs, followed by Sigmoid activation employed to normalize the attention weights. In the last, a channel-wise multiplication is performed with the original tensor.

The global average-pooling operation in the proposed spatial attention module aggregates spatial information across the channel. It creates a spatial attention map by leveraging the inter-spatial relationships between feature maps. A layer normalization (LN) is also applied after the global average-pooling operation, followed by Sigmoid activation employed to normalize the attention weights. In the last, a spatial-wise multiplication is performed with the original tensor. In short, the proposed channel attention is illustrated as Eq. (1), and spatial attention is illustrated as Eq. (2).

$$CA(x) = x * \sigma(LN(Gac(x))), \tag{1}$$

$$SA(x) = x * \sigma(LN(Gas(x))), \tag{2}$$

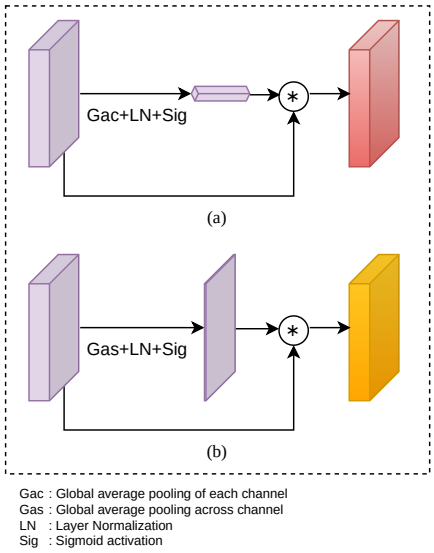

Fig. 2. The proposed channel attention module (a) and spatial attention module (b).

where Gac is the global average-pooling operation to summarize the spatial information from each channel, Gas is the global average-pooling operation to summarize the spatial information across the channel, x is an input of the attention module, LN indicates the layer normalization procedure, and σ refers to the Sigmoid activation function. The Sigmoid activation equation is represented as Eq. (3).

$$S(x) = \frac{1}{1 + e^{-x}}, \tag{3}$$

where x is an input of the function representing a logit score from the output of the network's last layer, and e is Euler's number.

Unlike [17], we do not apply a shared network containing a multi-layer perceptron (MLP) and a convolution operation after the global average-pooling operation in the channel and spatial attention modules to make the modules more efficient. Fig. 2 shows the proposed attention modules. The channel attention module is placed after the second and the third convolution layer. It will improve the quality of the middle and high-level features of the proposed model. The spatial attention module is only placed after the second channel attention module, which makes the model focus on 'where' is an informative spatial region after the proposed model extracts the high-level features of the proposed model. This placement also aims to make the model more efficient than placing these two attention modules after all convolution layers.

2.3 Classification Module

The classification module is used to classify features resulting from the feature extraction process on the backbone. It applies two dense or generally called fully-connected layers (FC). The first FC comprises 64 units, followed by batch normalization and ReLU (Rectified Linear Unit) activation layer. The final FC contains two units, followed by the Sigmoid activation, which transforms the output of the preceding layer into two possibilities expressing the prediction outcome as a first-class (male) or second-class (female). A dropout technique is used before the final fully-connected layer to stave off overfitting.

2.4 Face Detector

Face detection is conducted to detect and obtain a Region of Interest (RoI) of a face. It is obtained as a prior operation executed before the facial gender prediction process. It necessitates an efficient face detector as a partner that makes the facial gender detector appropriate to run in real-time. This work utilizes a face detector on [19]. This detector consists of twelve convolutional layers and six kinds of anchors, which only produce a few parameters. It drives the detector qualified to sprint in real-time. The RoI resulting from this detector will then be cropped and scaled to a particular size following the gender recognition input.

3. IMPLEMENTATION SETUP

In this work, the NVIDIA Tesla V100-PCIe 32GB is utilized as an accelerator to train the proposed CNN model. It uses UTKFace, LFW, and Adience Benchmark datasets as a training and validation process. Three hundred epochs are applied for the training process, which sets 10^{-2} as an initial learning rate. It performs a reducing learning rate mechanism in which the learning rate will decrease to 75% in every 20 epochs when there is no revision. In this work, the Adam optimizer is selected to revise the weight based on the Binary Cross-Entropy loss. A batch size of 256 is also chosen to hasten up the parallelism process of high-performance GPUs. Moreover, the proposed model is tested on Intel Core i7-9750H CPU @ 2.60GHz with 20GB RAM and implemented on Tensorflow and Keras framework.

4. EXPERIMENTAL RESULTS

This section describes the performance evaluation of the proposed CNN model on three datasets benchmark: UTKFace, LFW, and Adience. This section also compares the speed of the proposed model on a CPU device to the other CNN models.

4.1 Evaluation on Datasets

4.1.1 UTKFace

The dataset contains more than 23,000 face images labeled in age, gender, and ethnicity. For age, this dataset ranges from 0 to 116. The dataset also provides variations in resolution, expression, lighting, position, etc. The aligned and cropped faces version of the UTKFace dataset is used in this case. In this work, we divide the dataset into 70% (16,600 images) as training and 30% (7,108 images) as testing sets. A random permutation

Table 1. Evaluation results on UTKFace, LFW, and Adience datasets.

Model	Number of Parameters	Validation Accuracy (%)
Evaluation on UTKFace		
InceptionV3	21,806,882	88.26
SqueezeNet + BN	735,306	89.24
VGG13 + BN	34,467,906	89.28
VGG16 + BN	39,782,722	89.30
ResNet50V2	23,568,898	89.35
VGG11 + BN	34,413,698	89.43
Hamdi & Moussaoui [9]	530,034	89.97
MobileNet V2	2,260,546	90.49
Krishnan et al. (VGG-19) [20]	143,667,240	91.50
Krishnan et al. (ResNet-50) [20]	25,636,712	91.60
Krishnan et al. (VGG-16) [20]	138,357,544	91.90
Savchenko [21]	3,491,521	91.95
MPConvNet [10]	659,650	92.32
MudaNet	**674,760**	**92.66**
Evaluation on LFW		
Althnian et al. [22]	15,473,190	72.50
Rouhsedaghat et al. [23]	16,900	94.63
Greco et al. [24]	3,538,984	**98.73**
MudaNet	**674,760**	**96.22**
Evaluation on Adience Benchmark		
Althnian et al. [22]	15,473,190	83.30
Greco et al. [24]	3,538,984	84.48
Opu et al. [25]	210,050	**85.77**
MudaNet	**674,760**	**84.85**

split mechanism is set, which will randomly reorder the images in a different order than the earliest or previous order. By employing only 674,760 parameters, the MudaNet reaches 92.66% in validation accuracy. The result exceeds impressive CNN models such as VGG, ResNet, Inception, MobileNet, and SqueezeNet, as seen in Table 1. Furthermore, MudaNet reaches the validation accuracy overtaking the three lightweight CNN models, [9], SqueezeNet with batch normalization, and MPConvNet, which differed by 2.69, 2.17, and 0.34, respectively.

4.1.2 Labeled Faces in the Wild (LFW)

The dataset contains more than 13,000 face images with a low balance between females and males, about 23% and 77%, respectively. By applying a random permutation split, we divide the dataset into 70% (9,263 images) as training and 30% (3,971 images) as testing sets. With only 674,760 parameters, the MudaNet reaches 96,22% in validation accuracy. It also reaches competitive performance of validation accuracy with the two lightweight CNN models, [24] and [23], as seen in Table 1. However, the MudaNet becomes the second-best, after [24] with 3,538,984 parameters, which differed only by 2.51. Even so, the MudaNet has 80.93% fewer parameters.

Table 2. Comparison of model speeds on a CPU.

Model	Gender Recognition (FPS)	Face Detection + Gender Recognition (FPS)
InceptionV3	27.85	25.02
ResNet50V2	33.43	29.44
VGG16 + BN	38.36	33.07
VGG13 + BN	47.43	39.72
VGG11 + BN	52.66	43.66
MobileNet V2	55.94	46.16
Squeezenet + BN	97.05	71.08
MudaNet	**60.03**	**49.19**

4.1.3 Adience Benchmark

The dataset contains more than 26,000 face images labeled in gender and age. For age, this dataset ranges from 0 to 60. The dataset also provides variations in noise, appearance, pose, lighting, position, etc. This work eliminates data that contains missing values on the dataset, which generates 17,492 face images. By applying a random permutation split, we divide the dataset into 70% (12,244 images) as training and 30% (5,248 images) as testing sets. With only 674,760 parameters, the MudaNet reaches 84,85% in validation accuracy. It also reaches competitive performance of validation accuracy with the two lightweight CNN models, [25] and [24], as seen in Table 1. However, the MudaNet becomes the second-best, after [25] with 210,050 parameters, which differed only by 0.92.

4.2 Runtime Efficiency

MudaNet, a proposed CNN model with a few parameters, is offered to support advertisement displays that can perform on a CPU device in real-time. MudaNet produces only 674,760 parameters that can appropriately recognize facial gender in real-time, especially when it is incorporated with face detection. MudaNet can run 60.03 frames per second for gender recognition and 49.19 frames per second for gender detection, which is incorporated with face detection on [19]. As can be seen in

(a)

(b)

Fig. 3. The precise detection result (a) and the imprecise detection results (b) of the FG-CPU detector.

Table 2, MudaNet becomes the second-best rapid facial gender detector on a CPU compared to other common and lightweight CNN models. Even though Squeezenet has become the fastest facial gender detector on the CPU, the accuracy is lower than MudaNet, with a difference of 3.42. The recognition results of the FG-CPU detector can be seen in Fig. 3 (a). The white bounding box refers to a female face, and the blue bounding box refers to a male face.

4.3 Multi-pose Limitation

The FG-CPU detector with the MudaNet model is trained on the UTKFace dataset, which consists of some pose variations. However, the dataset does not provide numerous examples of every pose variation, particularly yaw and roll pose face. It makes the detector perform imprecise in recognizing some facial gender cases with the yaw and roll pose face. The inaccurate recognition results of the FG-CPU with the yaw and roll pose case can be seen in Fig. 3 (b).

5. CONCLUSION

This work presents a facial gender detector on a CPU (FG-CPU) with a lightweight CNN model. It proposes a multi-dilated convolution with attention modules (MudaNet) that utilize three-branch convolution layers. It constructs an efficient model that produces few parameters. The channel and spatial attention modules are employed to improve the previous convolution layer output quality. Based on the experimental results, MudaNet is proven to compete with other general and lightweight CNN models based on accuracy in three benchmark datasets: UTKFace, LFW, and Adience. As a result, the FG-CPU can operate 49.19 frames per second in real-time on a CPU device to recognize the facial gender. The proposed detector speed competes with other common and lightweight competitors' CNN models. In future work, a facial gender dataset with various poses will be explored to address the imprecise recognition of the multi-pose face.

ACKNOWLEDGEMENT

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] A. Greco, A. Saggese, and M. Vento, "Digital signage by real-time gender recognition from face images," in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*. IEEE, 2020, pp. 309–313.

[2] M. Alhalabi, N. Hussein, E. Khan, O. Habash, J. Yousaf, and M. Ghazal, "Sustainable smart advertisement display using deep age and gender recognition," in *2021 International Conference on Deci-*

sion Aid Sciences and Application (DASA). IEEE, 2021, pp. 33–37.

[3] C.-Y. Hsu, L.-E. Lin, and C. H. Lin, "Age and gender recognition with random occluded data augmentation on facial images," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 631–11 653, 2021.

[4] A. Priadana, M. R. Maarif, and M. Habibi, "Gender prediction for instagram user profiling using deep learning," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, 2020, pp. 432–436.

[5] M. A. Moreno-Armendáriz, H. Calvo, C. A. Duchanoy, A. Lara-Cázares, E. Ramos-Diaz, and V. L. Morales-Flores, "Deep-learning-based adaptive advertising with augmented reality," *Sensors*, vol. 22, no. 1, p. 63, 2021.

[6] K. Mishima, T. Sakurada, and Y. Hagiwara, "Low-cost managed digital signage system with signage device using small-sized and low-cost information device," in *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2017, pp. 573–575.

[7] Y. Bandung, Y. F. Hendra, and L. B. Subekti, "Design and implementation of digital signage system based on raspberry pi 2 for e-tourism in indonesia," in *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, 2015, pp. 1–6.

[8] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[9] S. Hamdi and A. Moussaoui, "Comparative study between machine and deep learning methods for age, gender and ethnicity identification," in *2020 4th International Symposium on Informatics and its Applications (ISIA)*. IEEE, 2020, pp. 1–6.

[10] 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.

[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] 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.

[13] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on information forensics and security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[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] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[16] 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.

[17] 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.

[18] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[19] 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.

[20] A. Krishnan, A. Almadan, and A. Rattani, "Understanding fairness of gender classification algorithms across gender-race groups," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 1028–1035.

[21] A. V. Savchenko, "Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output convnet," *PeerJ Computer Science*, vol. 5, p. e197, 2019.

[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] M. N. I. Opu, T. K. Koly, A. Das, and A. Dey, "A lightweight deep convolutional neural network model for real-time age and gender prediction," in *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*. IEEE, 2020, pp. 1–6.