A Fast Real-time Face Gender Detector on CPU using Superficial Network with Attention Modules

Adri Priadana, Muhamad Dwisnanto Putro, Changhyun Jeong 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, chjeong@mail.ulsan.ac.kr, acejo@ulsan.ac.kr

***Abstract*—A gender detector has become an essential part of digital signage to support the decision on providing relevant ads for each audience. Application installed in digital signage must be capable of running on low-cost or CPU devices to minimize system costs. This study proposed a fast face gender detector (Gender-CPU) that can sprint in real-time on CPU devices implemented on digital signage. The proposed architecture contains a superficial network with attention modules (SufiaNet). This architecture only consists of three convolution layers, making it super shallow and generating skimpy parameters. In order to redeem the lack of a super shallow network, the global attention module is assigned to improve the quality of the feature map resulting from the previous convolution layers. In the experiment, the training and validation process is conducted on the UTKFace, the Adience Benchmark, and the Labeled Faces in the Wild (LFW) datasets. The SufiaNet gains competitive accuracy compared to other common and light architectures on the three datasets. Moreover, the detector can run 84.97 frames per second on a CPU device, which is fast to run in real-time.**

***Index Terms*—face gender detector, real-time detector, digital signage, convolutional neural network, attention module**

I. Introduction

Digital content has been extensively grown in the past few years. It creates new opportunities for digital marketers, especially advertisers, to appeal the prospective customers through many digital platforms. The digital signage system is one of them that has become ubiquitous in many public areas where crowds assemble, such as supermarkets, airports, and hotels [1], [2]. It utilizes a screen board displaying varied content such as financial, saleable, and entertainment information. This platform has become an essential channel for offline advertising in modern cities [3].

Digital signage is an advertising platform that provides the dynamic customization of the ad contents following the audience looking at the screen [4]. Gender is one of the basic yet essential information of a face that can be used to segment the audience [5]. Recognizing gender can be used as the basis for advertising platforms to provide relevant ads for each audience [6]. It will establish more targeted advertising.

Recognizing gender is conducted by detecting and analyzing the face of the audience through a camera. This process is required to operate in real-time while the audiences face the platform. Moreover, digital signage requires a low-cost device to reduce the system cost [7], [8]. This issue creates an additional challenge if it implements in real-time. Therefore, this platform requires a gender recognition technology that can be implemented on a CPU or low-cost device.

Presently, the Convolutional Neural Network (CNN) has proved a lot of success in recognition works. Many researchers have developed diverse CNN architecture to build a recognition system, especially for face gender recognition. Hamdi and Moussaoui [9] proposed CNN architecture for gender prediction compared with some techniques such as Support Vector Machine and Random Forest. The proposed architecture achieved 89.97% accuracy on the UTKFace dataset. Ranjan et al. [10] used ResNet-101 architecture as a baseline to build new CNN architecture, namely HyperFace-ResNet, for gender prediction. In this architecture, the lower and deeper layers of ResNet are fused using an element-wise addition operator. This proposed architecture achieved 94% and 98% accuracy on the LFW and CelebA datasets, respectively. A CNN-based gender detector applied to digital signage also has been proposed in previous work. Greco et al. [4] employed MobileNetV2 architecture to design a gender detector applied in digital signage in real-time on an ARM-based CPU. This proposed detector achieved 95% accuracy. Although the detector used a tiny version of the CNN architecture, it generated 3,5 million numbers. The fewer parameters will increase the efficiency of the face gender detector, which makes it run fast. This study presents a real-time detector with a skimpy parameter that can detect a gender fast.

A fast CPU real-time face gender detector (Gender-CPU) proposed a superficial and lightweight architecture with attention modules (SufiaNet). The attention module is applied to improve the quality of the feature map by considering the interrelationship between channels. It generates few parameters and leads the detector to run fast. Therefore, this architecture can be applied on a low-cost device or CPU-based. The contribution of this work summarizes as follows:

1) A superficial and lightweight architecture with attention modules (SufiaNet) is proffered that produces few parameters. The attention modules can reinforce the object's important feature that impacts the accuracy of the recognition result.
2) A fast face gender detector that can run fast on a CPU device. The performance output gain competitive accuracy comparing to other architectures on UTKFace [11], Adience Benchmark [12], and Labeled Faces in the

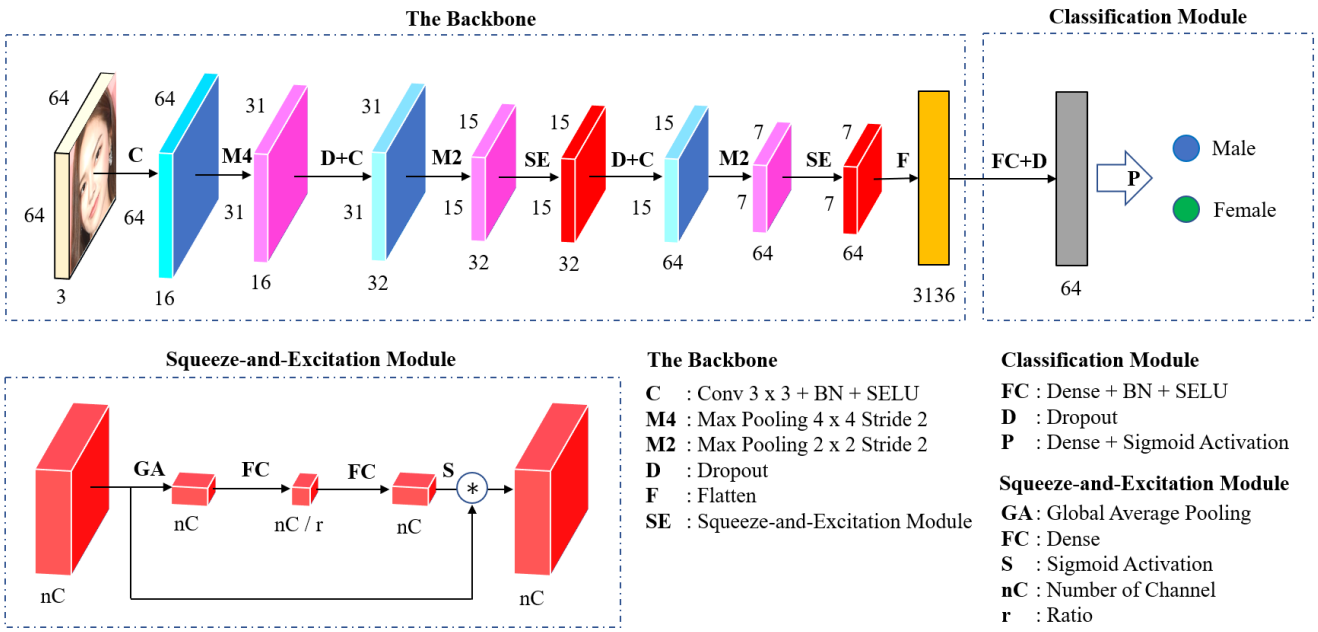

Fig. 1. The proposed architecture of the face gender detector. It uses a superficial and lightweight backbone with global attention modules to extract gender face features rapidly.

Wild (LFW) [13] datasets.

II. Proposed Architecture

The proposed architecture employs a sequence of convolution layers with global attention modules, as shown in Fig. 1. In this architecture, a backbone that consists of sequential convolution layers efficiently extracts gender features of faces. Hereafter, the classification module that consists of dense layers and activation functions predicts the gender of the face. The proposed architecture of this work produces 226,574 parameters.

A. The Backbone

The backbone module extracts gender features of faces using sequential convolution layers. The Gender-CPU proposes a superficial and lightweight architecture combined with a global attention module. This backbone only consists of three convolution layers with 3 × 3 kernel size. These layers are organized sequentially with a two-times increase in the number of kernels from 16, 32, and 64. It aims to obtain more information at a higher level of the feature extraction layer. We determine a small number of convolution layers to suppress the number of parameters that makes the architecture more efficient. In order to deal with the gradient problem, a batch normalization technique [14] and Scaled Exponential Linear Units (SELU) [15] activation are used after convolution operations. The dropout technique is also used before the second and third convolution layer to prevent overfitting [16].

Three max-pooling layers with two strides and different sizes are also applied to shrink the feature map. It is used to summarize the most essential features with high activation values [17]. A max-pooling layer with the 4 × 4 sizes is put after the first convolution layer. Further, a max-pooling layer with the 2 × 2 sizes is put after the second and the third convolution layer. The use of the 4 × 4 sizes of max-pooling in the first aims to summarize the broader area in the low-level feature.

The use of the few convolution layers makes this architecture super shallow and generates skimpy parameters. In order to redeem the lack of a super shallow network, the global attention module is assigned to enhance the quality of the feature map resulting from the previous convolution layers. We utilize the global attention mechanism, namely Squeeze-and-Excitation module (SE) [18], which performs a global average-pooling operation to aggregate each feature map resulting from the previous convolution layers. It produces a feature vector in which each value represents the features resume for the corresponding channel. After this operation, two sequential fully-connected layers are applied to capture the channel-wise dependencies. As it befits a bottleneck mechanism, the first layer is used to decrease the number of channels. The second layer is used to increase the number of channels. It is also used to match the original shape before performing a weighted summation with the original tensor resulting from the previous convolution layers. The global attention module function is described as:

$$CA(x) = x * \sigma(W_{FC2}(\delta(W_{FC1}(GA(x))))) \quad (1)$$

where x is an input of the SE module, GA is the global average pooling operation, σ indicates the Sigmoid function used to normalize the attention weights, δ refers to the ReLU

(Rectified Linear Unit) activation function, and W_{FC1} and W_{FC2} are learnable parameters in the two fully-connected layers. The global attention module is only located after the second and the third max-pooling to improve the quality of the feature map at the middle and high-level features of this architecture.

B. Classification Module

The classification module is used to calculate the probability of the gender class in order to predict the gender of the face. It consists of two fully-connected layers. The first layer consist of 64 units, batch normalization, and SELU activation function. It uses the SELU activation function to pass not only the positive values but also the negative values to avoid the loss of valuable information and dead neurons during activation [19]. The second layer consists of the Sigmoid activation function that renders the previous layer's output to possibilities representing the prediction result as a male or female. The equation of the Sigmoid activation function is described as:

$$S(x) = \frac{1}{1 + e^{-1}} \tag{2}$$

where x is a logit score from the neural network, and e is Euler's number. It also uses dropout operations before the second fully-connected layer to prevent overfitting.

C. Face Detector

Face detection is used as a preliminary process performed before face gender prediction. It serves to detect and get the face area as a Region of Interest (RoI). It required a face detector with efficient performance, especially to perform in real-time scenarios. In order to counter this issue, the LWFCPU [20] face detector is used in this work. This detector uses six types of anchors with only twelve convolutional layers. It only generates a few parameters. It makes the face detector capable of running fast on CPU or low computing devices in real-time detection. In advance of the RoI being entered into the gender recognition process, it will be cropped and scaled according to the predefined size of the gender recognition input.

III. Implementation Setup

The proposed architecture is trained on the NVIDIA Tesla V100-PCIe 32GB as an accelerator. Further, it is tested on Intel Core i7-9750H CPU @ 2.60GHz with 20GB RAM. The UTKFace, Adience Benchmark, and LFW datasets are used in the training and validation process. The total epoch is 300 in the training stage with 10^{-2} as an initial learning rate. In this implementation, the reducing learning rate mechanism is applied. During the training process, the learning rate will reduce to 75% when there is no improvement in every 20 epochs. The Adam is set as an optimizer to update the weight according to Binary Cross-Entropy loss. In order to computational speed up from the parallelism of high-performance GPUs, a batch size of 256 is applied. The proposed architecture is implemented on Keras 2.3.1 and the Tensorflow 2.0 framework.

TABLE I
Evaluation results on UTKFace, Adience, and LFW datasets

Architecture	Number of Parameters	Validation Accuracy (%)
Evaluation on UTKFace (Aligned and Cropped Faces)		
VGG16 + Batch Normalization	39,782,722	89.30
VGG13 + Batch Normalization	34,467,906	89.28
VGG11 + Batch Normalization	34,413,698	89.43
ResNet50V2	23,568,898	89.35
InceptionV3	21,806,882	88.26
SqueezeNet + Batch Normalization	735,306	89.24
Hamdi & Moussaoui [9]	530,034	89.97
MobileNet V2	2,260,546	90.49
Krishnan et al. (VGG-19) ([21]	143,667,240	91.50
Krishnan et al. (ResNet-50) [21]	25,636,712	91.60
Krishnan et al. (VGG16) [21]	138,357,544	91.90
Savchenko [22]	3,491,521	91.95
SufiaNet	**226,574**	**92.05**
Evaluation on Adience Benchmark		
Althnian et al. [23]	15,473,190	83.30
Greco et al. [24]	3,538,984	84.48
Opu et al. [25]	210,050	85.77
SufiaNet	**226,574**	**84.60**
Evaluation on LFW		
Althnian et al. [23]	15,473,190	72.50
Rouhsedaghat et al. [26]	16,900	94.63
Greco et al. [24]	3,538,984	98.73
SufiaNet	**226,574**	**95.66**

IV. Experimental Results

In this experiment, three datasets are used to measure the performance evaluation of gender prediction. This section describes the examination result of the proposed architecture on the datasets benchmark. This section also investigates the speed of the SufiaNet on a CPU and compares it to other architectures.

A. Evaluation on Datasets

1) UTKFace (Aligned and Cropped Faces): The proposed architecture is evaluated in the UTKFace dataset to test the performance of the face gender detector. The dataset consists of 23,708 face images ranging from 0 to 116. It is labeled in gender, age, and ethnicity. The dataset also covers huge variations such as illumination, expression, pose, resolution, etc. In this experiment, the dataset is divided into 70% as training and 30% as testing sets with a random permutation split. The 16,600 images are used as training and the 7,108 images as testing sets. As a result, the SufiaNet achieves 92,06% in validation accuracy with only 226,574 parameters. The result exceeds some outstanding architecture such as Inception, ResNet, VGG, and MobileNet. Furthermore, SufiaNet achieves the validation accuracy surpassing the two light architectures, SqueezeNet with batch normalization and [9], which differed by 2.81 and 2,08, respectively, as can be seen in Table I.

2) Adience Benchmark: In the second evaluation, the Adience Benchmark dataset is used to test the performance of the face gender detector. The dataset consists of 26,580 face images covering age variations ranging from 0 to 60. It is labeled in gender and age. The dataset also covers huge

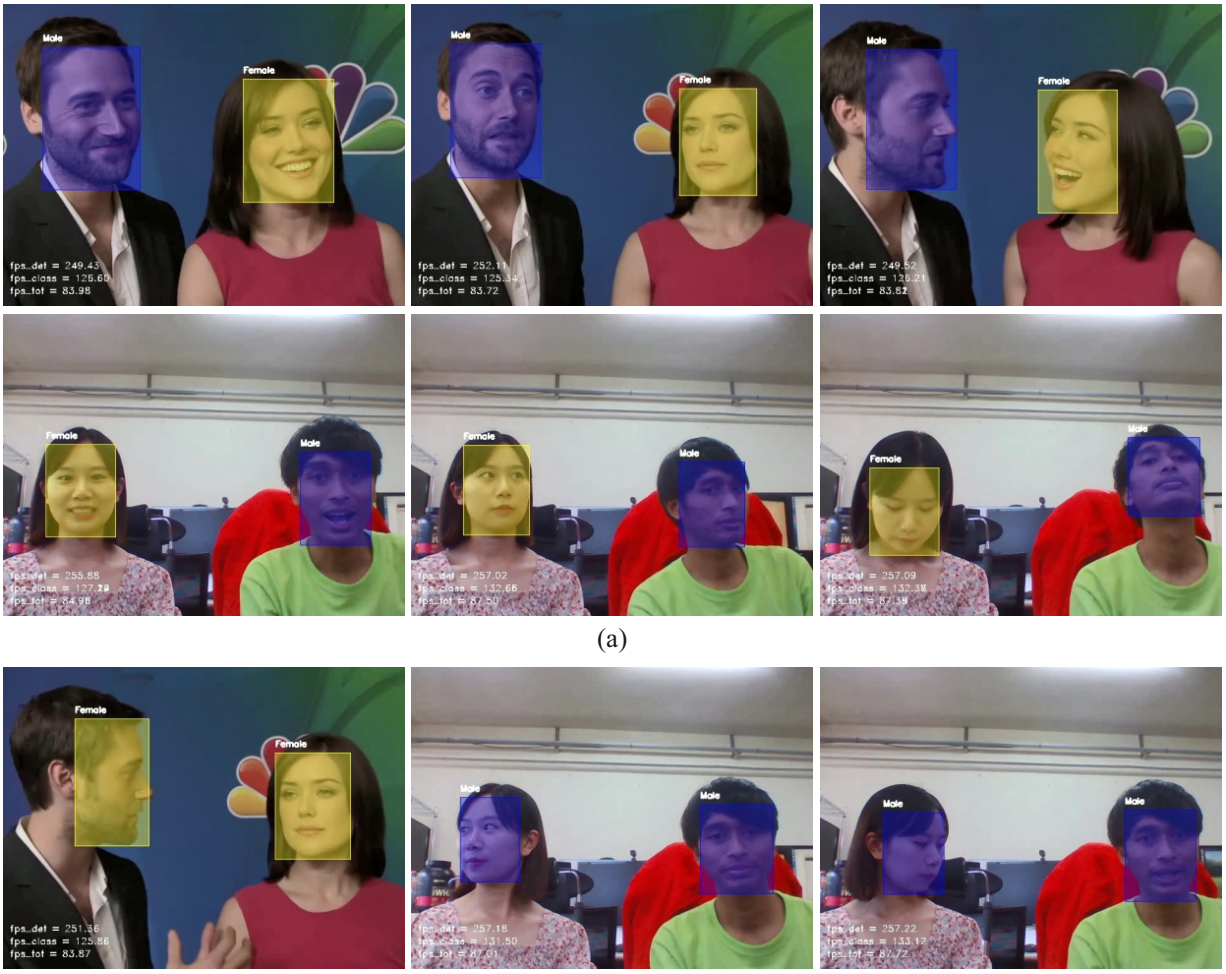

Fig. 2. The correct prediction result (a) and the incorrect prediction results (b) of the Gender-CPU detector.

variations such as lighting, pose, appearance, noise, etc. In this experiment, some pre-processing is conducted, such as eliminating data that contains missing values on the dataset. It produces 17,492 face images. With a random permutation split, the dataset is divided into 70% as training and 30% as testing sets. The 12,244 images are used as training and the 5,248 as testing sets. As a result, the SufiaNet achieves 84,60% in validation accuracy with only 226,574 parameters. The SufiaNet achieves competitive performance concerning validation accuracy of the two light architectures such as [25] and [24]. However, the proposed detector occupies the second-best, below the performance of [25] with 210,050 parameters, which differed by 1.17, as can be seen in Table I.

3) Labeled Faces in the Wild (LFW): In the third evaluation, the LFW dataset is used to test the performance of the face gender detector. The dataset consists of 13,234 face images with a significant imbalance between males (77%) and females (23%). With a random permutation split, the dataset is divided into 70% as training and 30% as testing sets. The 9,263 images are used as training and the 3,971 as testing sets. As a result, the SufiaNet achieves 95,66% in validation accuracy with only 226,574 parameters. The SufiaNet achieves competitive performance concerning validation accuracy of the two light architectures such as [24] and [26]. However, the proposed detector occupies the second best. It is below the performance of [24] with 3,538,984 parameters, which differed by 3.17, as can be seen in Table I. Even so, the SufiaNet has 93.6% fewer parameters.

B. Runtime Efficiency

The SufiaNet, with a few parameters, is specially designed to be implemented in real-time on CPU-based supporting the digital signage. The proposed architecture generates only 226,574 parameters. Therefore, it can be efficient when integrated with face detection to recognize the face gender in real-time. The proposed detector achieves 127.50 FPS for gender recognition and 84.97 FPS for the integration between the face detection using LWFCPU [20] and the proposed gender recognition. It makes our proposed detector the fastest on a CPU compared to other common and light architectures shown

TABLE II
COMPARISON OF ARCHITECTURE SPEEDS ON CPU

Architecture	Gender Recognition (FPS)	Face Detection + Gender Recognition (FPS)
InceptionV3	31.85	28.35
ResNet50V2	35.54	31.24
VGG16 + Batch Normalization	40.32	34.91
VGG13 + Batch Normalization	47.47	40.27
VGG11 + Batch Normalization	51.83	43.21
MobileNet V2	57.42	47.59
Squeezenet + Batch Normalization	95.18	69.27
SufiaNet	**127.50**	**84.97**

in Table II. Fig. 2 (a) shows the correct prediction results of the Gender-CPU detector on the CPU. The blue color indicates a male face, and the yellow bounding box indicates a female face.

C. Limitation

The Gender-CPU detector with SufiaNet architecture for gender prediction is training on the UTKFace dataset using the aligned and cropped faces version that covers variations in the pose. However, it does not have many instances, especially face in full yaw pose. It causes the detector resulting an incorrect prediction in a few cases when it predicts the gender with the face in full yaw pose. Fig. 2 (b) shows the false prediction outcomes of the detector with the yaw pose case.

V. CONCLUSION

This study proposes a fast real-time face gender detector with light architecture. It offers a superficial network with attention modules (SufiaNet) that only consists of three convolution layers. It makes the architecture super shallow and generates skimpy parameters. The global attention module is utilized to redeem the lack of a super shallow network. It improves the quality of the feature map resulting from the previous convolution layers. The SufiaNet gained competitive accuracy compared to other common and light architectures on the UTKFace, the Adience Benchmark, and the Labeled Faces in the Wild (LFW) datasets. As a result, the Gender-CPU can run 84.97 frames per second in recognizing the gender of the face when working on CPU devices in real-time. The speed of the proposed detector outperforms other common and light competitors' architecture. In future work, the novel attention module with minimal operation and parameters can be designed to increase the detector's speed.

REFERENCES

[1] M. Garaus, U. Wagner, and R. C. Rainer, "Emotional targeting using digital signage systems and facial recognition at the point-of-sale," *Journal of Business Research*, vol. 131, pp. 747–762, 2021.

[2] H. Okada, S. Sato, T. Wada, K. Kobayashi, and M. Katayama, "Preventing degradation of the quality of visual information in digital signage and image-sensor-based visible light communication systems," *IEEE Photonics Journal*, vol. 10, no. 3, pp. 1–9, 2018.

[3] Y. Park, H. Yang, T. Dinh, and Y. Kim, "Design and implementation of a container-based virtual client architecture for interactive digital signage systems," *International Journal of Distributed Sensor Networks*, vol. 13, no. 7, p. 1550147717717864, 2017.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[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] 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.

[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] 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.

[13] 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.

[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] 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.

[18] 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.

[19] H. Liu, J. Luo, B. Huang, X. Hu, Y. Sun, Y. Yang, N. Xu, and N. Zhou, "De-net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sensing*, vol. 11, no. 20, p. 2380, 2019.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[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.

[26] 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.