Human Face Detector with Gender Identification by Split-based Inception Block and Regulated Attention Module

Adri Priadana[1], Muhamad Dwisnanto Putro[2], Duy-Linh Nguyen[1], Xuan-Thuy Vo[1], and Kang-Hyun Jo[1]

[1] Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, South Korea
[2] Department of Electrical Engineering, Universitas Sam Ratulangi, Manado, Indonesia
priadana3202@mail.ulsan.ac.kr, dwisnantoputro@unsrat.ac.id, ndlinh301@mail.ulsan.ac.kr, xthuy@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

Abstract. Smart digital advertising platforms have been widely arising. These platforms require a human face detector with gender identification to assist them in the determination of providing relevant advertisements. The detector is also prosecuted to identify the gender of a masked face in post-coronavirus situations and demanded to operate on a CPU device to lower system expenses. This work presents a lightweight Convolution Neural Network (CNN) architecture to build a gender identification integrated with face detection to respond to these issues. This work proposes a split-based inception block to efficiently extract features at various sizes by partially applying different convolution kernel sizes, levels, and regulated attention module to improve the quality of the feature map. It produces slight parameters that drive the architecture efficiency and can operate quickly in real-time. To validate the performance of the proposed architecture, UTKFace and Labeled Faces in the Wild (LFW) datasets, modified with an artificial mask, are utilized as training and validation datasets. This offered architecture is compared to different lightweight and deep architectures. Regarding the experiment results, the proposed architecture outperforms masked face gender identification on the two datasets. In addition, the proposed architecture, which integrates with face detection to become a human face detector with gender identification can run 135 frames per second in real-time on a CPU configuration.

Keywords: Human Face Detector · Face Gender Identification · Convolutional Neural Network (CNN) · Split-based Inception Block · Regulated Attention Module.

1 Introduction

The advancement of information technology has stimulated the rapid development of smart digital advertising, not only in online media but also in offline media. It is proven because these platforms appear in many public places, such as airports, stations, and markets [4]. Practically, smart digital advertising platforms are handily personalized and customized. Therefore, it can display dynamic contents as determined by the provider. Nevertheless, the market demands effective mechanisms that make these platforms can provide targeted advertising [1]. This mechanisms will offer more advantages in the digital advertising strategy [20].

Providing targeted advertising can be accomplished by personalizing the audience facing the platform. The audience's gender, which is an essential attribute, can be used in segmenting the readers. These platforms can provide better appropriate advertising for each reader by recognizing their gender [15]. This scheme can be achieved with the reader's face detection and classification.

Nowadays, Convolutional Neural Network (CNN) has verified a bunch of victories in image-based detection and classification tasks. The common direction in designing CNN architectures is to develop deeper architectures to reach higher accuracy [7,25]. However, it tends to generate architecture with a large number of parameters. It makes the architecture inefficient to operate, especially on low-cost devices in real-time. In the case of advertising platform implementation, it requires a low-cost device, such as a CPU device, to minimize the implementation expense [3,13]. Hence, it requires an efficient face gender detector, which can be suitably operated on a CPU in real-time.

A new challenge arises after the spread of the COVID-19 virus extensively. It makes people required or used to wear masks on their faces when they are traveling. It makes part of the face area occluded, such as the mouth, which is one of the essential features for recognizing gender through the face. Therefore, it needs an efficient human face detector with gender identification ability that can detect and recognize the gender of a masked face. This work presents an efficient human face detector with gender identification by a few parameters that can efficiently detect and identify a masked face gender while maintaining its performance.

An efficient CPU-based human face detector with gender identification called GenderMask-CPU proposed a lightweight architecture with a split-based inception block and regulated attention module (SiramNet). The split-based inception block is offered to efficiently extract features at various sizes by partially applying different convolution kernel sizes and levels. The regulated attention module, which consist of the channel and spatial, are employed to enhance the feature map grade. It produces scant parameters and guides the detector to work efficiently and fast. In summary, the main contribution of this work is twofold, i.e.,

1. An efficient architecture with a split-based inception block and regulated attention module (SiramNet) is proposed, which generates slight parameters. The split-based inception block can efficiently extract multi-size feature areas of the feature maps. The attention module can maintain the essential features of the face area, which can increase the gender accuracy of the classification.

2. A fast human face detector with gender identification is introduced, which can operate in real-time on a CPU device efficiently and fast. The performance of the offered architecture is proven to compete with other deep and light CNN architectures on UTKFace [30] and Labeled Faces in the Wild (LFW) [10] datasets, modified with an artificial mask utilized from [2].

2 Related Work

In recent years, CNN architectures, designed for face gender recognition work, have progressed with impressive improvement, especially in performance. Various modified versions of CNNs have been developed to optimize face gender recognition. HyperFace-ResNet [21], a CNN architecture, was proposed to perform gender recognition from a face. The architecture develops and adjusts ResNet [5] architecture and reaches good performance on LFW datasets. In [4], a CNN architecture has been employed to recognize gender and implemented in the monitoring system. The architecture utilized MobilenetV2 [22] architecture and generated 3.5 million parameters.

Nowadays, efficient face gender detectors emerge specially designed for CPU devices to encounter market demand which can reduce implementation costs. MPConvNet [17] based on the CNN architecture was developed and generated 659,650 parameters. The architecture proposed a multi-perspective convolution used to capture various feature regions of the object. The architecture reaches good performance on UTKFace and LFW datasets. SufiaNet [16] based on the CNN architecture was developed and only generated 226,574 parameters. SufiaNet [16] is a shallow architecture supported by a global attention module. The architecture gains sufficient performance on UTKFace and LFW datasets.

3 The Proposed Method

This work proposes a CNN architecture to recognize a gender of a masked face, as shown in Fig. 1. This architecture is structured as a backbone and classification module, generating 441,460 parameters.

3.1 The Backbone

CNN-based feature extraction has shown excellent performance. However, this extractor tends to generate enormous parameters [6]. Therefore, an efficient backbone module is proposed to develop a fast architecture, especially one that can run on the CPU in real time. This architecture employs three main convolution layers with same 3×3 kernel size, managed sequentially by three times the number of kernels growing, i.e. 16, 32, and 64. This mechanism seeks to acquire more information on the latter layers. Following each convolution layer, a batch normalization (BN) method and Leaky ReLU (Leaky Rectified Linear Unit) activation are applied to deal with the vanishing gradient. A dropout strategy puts previous to the final convolution operation is also used to impede overfitting. Three max-pooling operations are assigned in this backbone to down-sample the feature maps. One layer of 4 × 4 and two layers of 2 × 2 sizes max-pooling with two strides are applied.

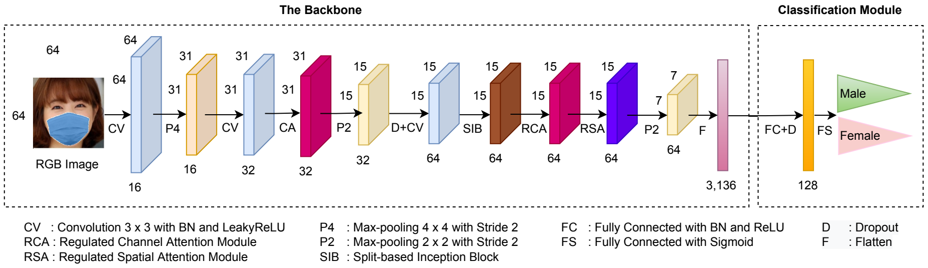

The Backbone
Classification Module
64
64
RGB Image
CV
64
64
16
P4
31
31
16
CV
31
31
32
CA
31
31
32
P2
15
15
32
D+CV
15
15
64
SIB
15
15
64
RCA
15
15
64
RSA
15
15
64
P2
7
7
64
F
3,136
FC+D
128
FS
Male
Female
CV : Convolution 3 x 3 with BN and LeakyReLU
RCA : Regulated Channel Attention Module
RSA : Regulated Spatial Attention Module
P4 : Max-pooling 4 x 4 with Stride 2
P2 : Max-pooling 2 x 2 with Stride 2
SIB : Split-based Inception Block
FC : Fully Connected with BN and ReLU
FS : Fully Connected with Sigmoid
D : Dropout
F : Flatten

Fig. 1. The proposed architecture of the gender identification of masked faces contains a backbone with a split-based inception block and regulated attention module.

3.2 The Split-based Inception Block

To improve the feature extractor on the backbone module, this work proposes a split-based inception block and applies the block after the last convolution layer. Inspired by the inception block [26], this module employs four branches of convolution layer with different levels and kernel sizes, as shown in Fig. 2. They are convolution layers with 1 × 1, 3 × 3, two times 3 × 3, and 5 × 5 kernel sizes. Unlike the original inception block that applies the convolution layer with the same number of a kernel as the input, this block divides the input feature map $\mathbf{X}$ become four components $[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}]$. Then, it applies convolution operation with different levels and kernel sizes mentioned before, which is represented as follows:

$$\begin{aligned} SIB(\mathbf{X}) = \mathbf{X} + (SELU(BN(C1(D(\mathbf{X_1})))) \oplus SELU(BN(C3(D(\mathbf{X_2})))) \\ \oplus SELU(BN(C3(D(SELU(BN(C3(D(\mathbf{X_3}))))))) \\ \oplus SELU(BN(C5(D(\mathbf{X_4})))), \end{aligned} \tag{1}$$

where $C1$, $C2$, $C3$ are convolution layers with 1 × 1, 3 × 3, and 5 × 5 kernel sizes, respectively. $SELU$ is Scaled Exponential Linear Units (SELU) activation [12], D is dropout operation, BN is batch normalization operation, and $\oplus$ is the concatenate operation. This block will extract more information from different levels and area sizes efficiently. At the last stage, a residual mechanism [6] is applied to combine the concatenate operation result with the input feature map $\mathbf{X}$ by an addition operation.

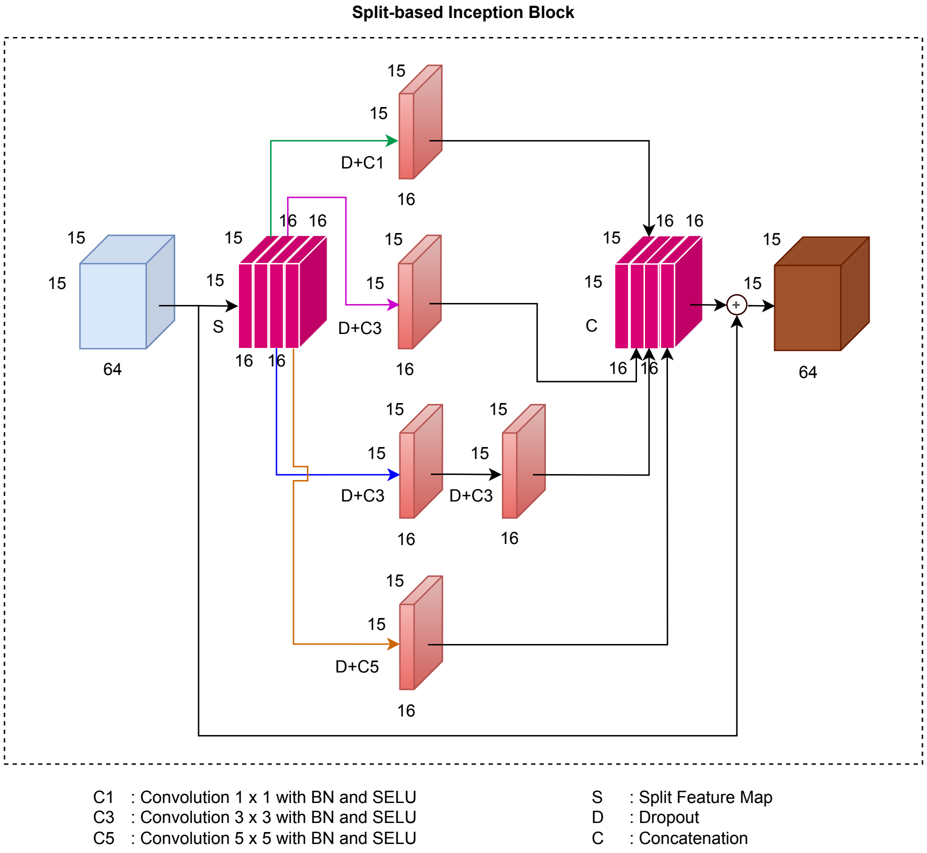

Split-based Inception Block
15
15
64
15
15
S
16 16
16 16
15
15
D+C1
16
15
15
D+C3
16
15
15
D+C3
16
15
15
D+C3
16
15
15
D+C5
16
15
15
C
16 16
16 16
15
15
64
C1 : Convolution 1 x 1 with BN and SELU
C3 : Convolution 3 x 3 with BN and SELU
C5 : Convolution 5 x 5 with BN and SELU
S : Split Feature Map
D : Dropout
C : Concatenation

Fig. 2. The proposed Split-based Inception block.

3.3 The Regulated Attention Module (RAM)

A backbone with few parameters feebly discriminates interest features of the face. Therefore, the Regulated Attention module (RAM) is proposed and applied to improve essential facial features. This module consists of a regulated channel attention module (RCA) and a regulated spatial attention module (RSA). Inspired by the attention module in [9], RCA performs a global average-pooling operation to aggregate each feature map based on channel. However, we do not use fully connected layers but apply softmax activation directly after the pooling operation to calculate the probability of channel importance level. A softmax activation is used rather than the sigmoid activation because it can establish long-range channel dependency [29]. Imbued from the attention module [18], this architecture puts a 3×3 depthwise convolution layer before the pooling operations to allow the individual channel to expand learning efficiently, as shown in Fig. 3. Different from [18], this architecture only applies global average pooling to squeeze the number of operations. Further, we proposed a 1×1 depthwise convolution layer located after softmax activation and before performing a channel-wise multiplication in the last step to regulate the attention weights individually represented as follows:

$$RCA(\mathbf{X}) = \mathbf{X} * DC1(\sigma(GA(DC3(\mathbf{X})))), \tag{2}$$

where $DC1$ and $DC3$ are 1×1 and 3×3 depthwise convolution layers, respectively. GA is a global average-pooling operation and σ is a softmax activation. $\mathbf{X}$ is an input of the RCA.

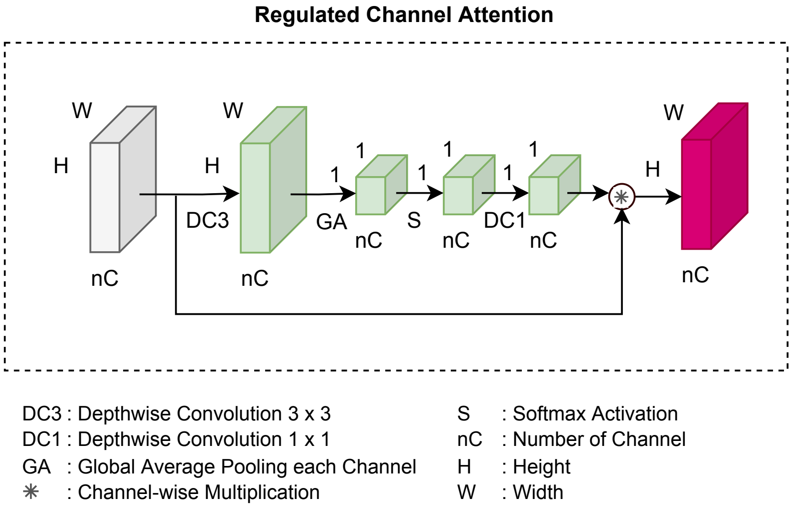

Regulated Channel Attention
W
H
nC
W
H
DC3
nC
1
1
GA
nC
1
1
S
nC
1
DC1
nC
W
H
nC
DC3 : Depthwise Convolution 3 x 3
DC1 : Depthwise Convolution 1 x 1
GA : Global Average Pooling each Channel
⁕ : Channel-wise Multiplication
S : Softmax Activation
nC : Number of Channel
H : Height
W : Width

Fig. 3. The proposed Regulated Channel Attention module.

Motivated by [28], a global average-pooling operation is assigned on RSA to aggregate spatial features across the channel. This operation renders a feature vector and describes the feature overview of the corresponding channel. However, a softmax activation is voted than a sigmoid activation to calculate the spatial importance level. This activation can establish spatial dependency. A 1×1 depthwise convolution layer is also applied after softmax activation and before performing a spatial-wise multiplication to regulate the attention weights with a shared parameter, as shown in Fig. 4. It is represented as follows:

$$RSA(\mathbf{X}) = \mathbf{X} * DC1(\sigma(GA(\mathbf{X}))), \tag{3}$$

where $DC1$ is a 1×1 depthwise convolution layer and GA is a global average-pooling across the channel. σ is a softmax activation and $\mathbf{X}$ is an input of RSA.

RCA is assigned following the second convolution layer and the split-based inception block. This module will enhance the grade of the intermediate and

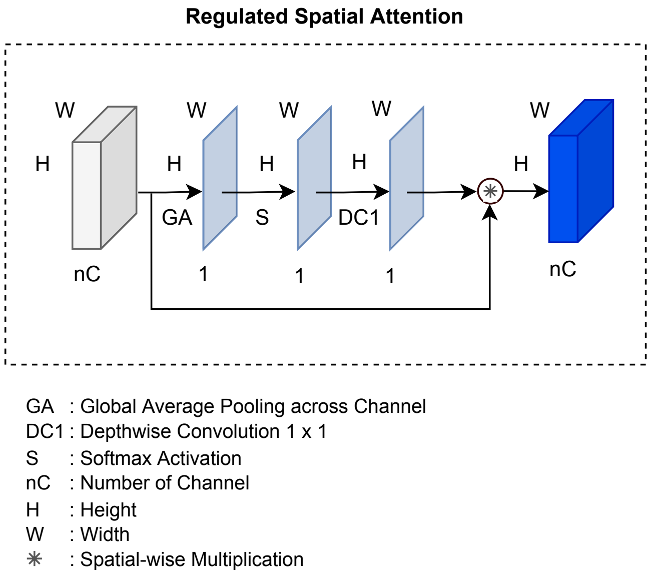

Regulated Spatial Attention
W
H
nC
W
H
GA
1
W
H
S
1
W
H
DC1
1
W
H
nC
GA : Global Average Pooling across Channel
DC1 : Depthwise Convolution 1 x 1
S : Softmax Activation
nC : Number of Channel
H : Height
W : Width
✳ : Spatial-wise Multiplication

Fig. 4. The proposed Regulated Spatial Attention module.

latter features. On the other hand, RSA is only assigned following the last RCA, which drives the architecture to focus on the location of informative spatial features after it extracts the high-level features.

3.4 Classification Module

The backbone module is tasked to extract features from masked faces. Then, the results will be fed to the classification module employed to reckon the probability of each gender class. This operation leads to deciding whether the masked face is male or female. This classification module is composed of two dense layers with 128 and 2 units, respectively. A batch normalization and ReLU (Rectified Linear Unit) activation are applied after the first dense layer, and the Sigmoid activation is applied after the second dense layer. The Sigmoid activation will render the input into scenarios that could describe the prediction decision of whether the masked face is male or female. In order to discourage overfitting, it applies a dropout operation after the ReLU activation.

3.5 Face Detector

In this work, face detection is required for integrating with masked face gender recognition to build a masked face gender detector. It is employed to locate and get the region of the face or masked face referred to as a Region of Interest (RoI). An efficient face detection model with cheap operation is required to operate brief in the real-time. Hence, a face detector named LWFCPU [19] is utilized. It employs only several convolutional layers that generate slight parameters. The RoI, which comes from the face detection operation, will become an input of the proposed gender recognition architecture. It will be resized and cropped to a particular size appropriate for the architecture input.

4 Experimental Settings

4.1 Dataset Pre-Processing

In this work, UTKFace and LFW datasets, which are labeled as females and males, are used for training and validation separately. Firstly, each facial image of these datasets is resized into 64×64 pixels appropriated with the input of the proposed gender recognition architecture. To generate masked face instances, we follow [2] to overlay one type of mask (Surgical) on UTKFace and LFW images, which produce 22,841 and 10,374 masked face images, respectively, and the examples are shown in Fig. 5. In this experiment, each dataset is split using a random permutation mechanism into 70% as a training set and 30% as a validation set. This mechanism will generate the unique order of the instances.

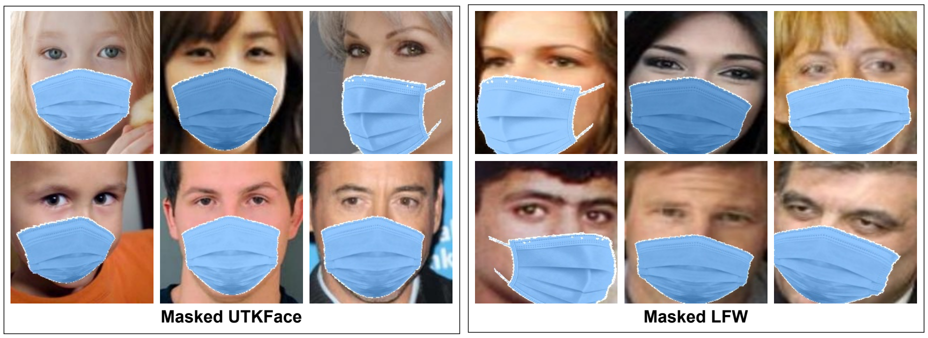


Fig. 5. The examples of the UTKFace and Labeled Faces in the Wild (LFW) datasets, modified with an artificial mask utilized from [2].

4.2 Implementation Details

The experiment is executed on the NVIDIA GTX 1080Ti 11GB to accelerate the training on the proposed architecture by using Tensorflow and Keras framework libraries. UTKFace and LFW datasets modified with an artificial mask referenced from [2] are used as training and validation to ratify the performance of the proposed architecture, which trains with three hundred epochs. The Adam optimizer is employed to optimize the weight on the Binary Cross-Entropy loss.

The datasets are trained by using 10^{-2} initial learning rate, which will reduce to 75% if the accuracy does not improve in every 20 epochs. Intel Core i7-9750H CPU@2.6 GHz with 20GB RAM is used to investigate the speed in frame per second (FPS) of the proposed architecture and the detector.

5 Results

5.1 Evaluation on Datasets

UTKFace. A face dataset labeled in gender, age, and ethnicity, is used for training and validation to ratify the performance of the proposed architecture. This dataset consists of 23,708 instances with various positions, expressions, resolutions, and lighting. This dataset also covers age variations ranging from 0 to 116. This dataset was modified with an artificial mask utilized from [2] and generated 22,841 masked face images. The proposed architecture, which only employs 441,460 parameters, gains 91.17% of validation accuracy. The proposed architecture outperforms deep CNN architectures [24,7,27], as sown in Table 1. Moreover, the proposed architecture reaches accuracy surpassing the three lightweight architectures, SqueezeNet [11], SufiaNet [16], and MPConvNet [17], which differed by 2.4, 1.16, and 0.98, respectively.

Table 1. Evaluation results on UTKFace dataset, modified with an artificial mask utilized from [2].

Architectures	Number of Parameters	Validation Accuracy
MobileNetV2 [23]	2,260,546	87.93
ResNet50V2 [7]	23,568,898	87.99
VGG13 [24] with BN	34,467,906	88.07
SqueezeNet [11] with BN	735,306	88.77
VGG16 [24] with BN	39,782,722	89.23
VGG11 [24] with BN	34,413,698	89.26
InceptionV3 [27]	21,806,882	89.64
SufiaNet [16]	226,574	90.01
MPConvNet [17]	659,650	90.19
SiramNet (ours)	**441,460**	**91.17**

LFW. A face dataset labeled in gender consists of 13,234 instances with unbalance proportion between males and females, about 77% and 23%. This dataset was also modified with an artificial mask utilized from [2] and generated 10,374 masked face images. The proposed architecture, which only employs 441,460 parameters, gains 95.64% of validation accuracy. The proposed architecture also outperforms deep CNN architectures [24,7,27], as sown in Table 2. Moreover, the proposed architecture also reaches accuracy surpassing the three lightweight architectures, SqueezeNet [11], SufiaNet [16], and MPConvNet [17], which differed by 1.38, 0.38, and 0.27, respectively.

Table 2. Evaluation results on LFW dataset, modified with an artificial mask utilized from [2].

Architectures	Number of Parameters	Validation Accuracy
MobileNetV2 [23]	2,260,546	79.93
VGG13 [24] with BN	34,467,906	91.18
InceptionV3 [27]	21,806,882	92.58
ResNet50V2 [7]	23,568,898	92.96
VGG16 [24] with BN	39,782,722	93.35
VGG11 [24] with BN	34,413,698	93.88
SqueezeNet [11] with BN	735,306	93.99
SufiaNet [16]	226,574	95.13
MPConvNet [17]	659,650	95.37
SiramNet (ours)	**441,460**	**95.64**

5.2 Ablation Study

This work performs the ablation study to investigate how much the proposed split-based inception block and attention module will impact the validation accuracy result. This ablative study conducts by repealing the block or module and then calculating the validation accuracy on the UTKFace dataset. As can be seen in Table 3, utilizing the proposed split-based inception block and applying this block after the last convolution layer can increase the accuracy by 0.12%. The proposed RCA can escalate the accuracy by 0.38%. Moreover, the proposed RSA module can also escalate the accuracy by 0.2% by adding only two parameters.

Table 3. Ablation study on UTKFace dataset, modified with an artificial mask utilized from [2].

Group Split Inception Block	Regulated Channel Attention Module	Regulated Spatial Attention Module	Number of Parameters	Validation Accuracy
			426,338	90.47
✓			440,306	90.59
✓	✓		441,458	90.97
✓	✓	✓	441,460	91.17

5.3 Runtime Efficiency

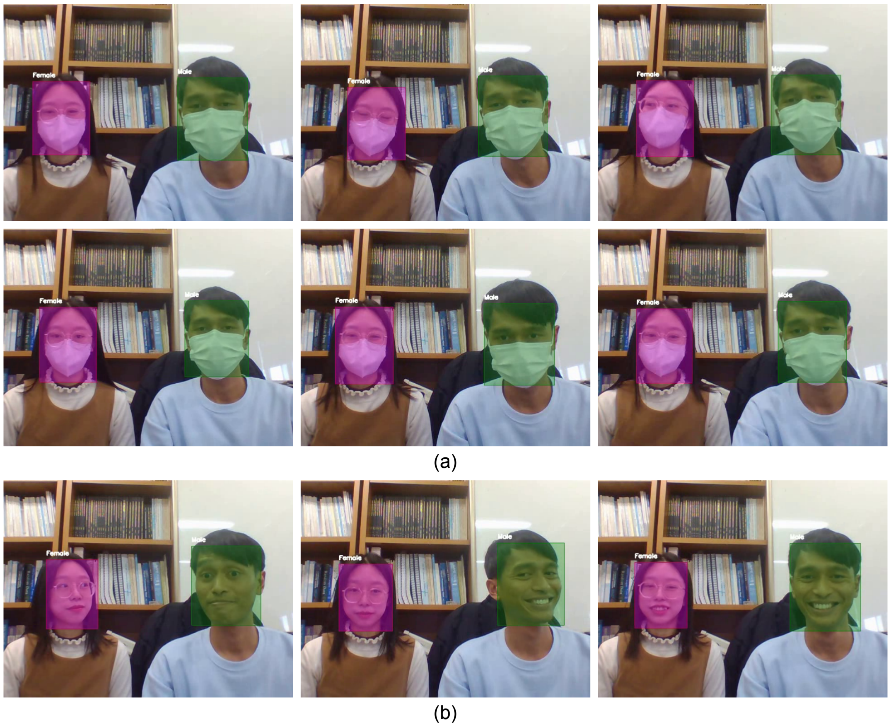
The proposed architecture recognizes gender from a masked face using only 441,460 parameters. The architecture operates in real-time at 272.80 and 135.02 frames per second for gender identification and gender identification integrated with face detection [19], respectively. The proposed efficient architecture becomes the second fastest compared to other deep and lightweight architectures, as shown in Table 4. Even though SufiaNet [16] has become the fastest architecture, the validation accuracy is not better than our proposed architecture, with a difference of 1.16 and 0.38 on the UTKFace and LFW datasets, modified with an artificial mask utilized from [2], respectively. Fig. 6 shows the recognition result of the GenderMask-CPU, in which the green bounding box means a male face and the magenta bounding box means a female face. Although this detector is specially designed for faces with mask shown in Fig. 6 (a), it can also work on faces without mask shown in Fig. 6 (b).

Table 4. Runtime efficiency on an Intel Core i7- 9750H CPU.

Architectures	**Number of Parameters**	**GFLOPs**	**Gender Recognition (FPS)**	**Gender Recognition integrated with Face Detection (FPS)**
VGG16 [24] with BN	39,782,722	2.2900	43.28	37.15
VGG13 [24] with BN	34,467,906	1.6100	51.14	42.76
VGG11 [24] with BN	34,413,698	1.2700	55.49	45.71
ResNet50V2 [7]	23,568,898	0.5710	57.19	46.79
InceptionV3 [27]	21,806,882	0.4050	64.43	51.47
MobileNetV2 [23]	2,260,546	0.0501	118.71	81.20
SquezeeNet [11] with BN	735,306	0.0833	231.40	122.75
MPConvNet [17]	659,650	0.0670	269.86	132.04
SufiaNet [16]	226,574	0.0218	327.29	145.37
SiramNet (ours)	**441,460**	**0.0293**	**272.80**	**135.02**

5.4 Attention Modules Comparison

The proposed regulated attention module (RAM) is also compared with other common attention modules, as shown in Table 5. This module compares with Squeeze-and-Excitation (SE) [8], Bottleneck Attention Module (BAM) [14], and Convolutional Block Attention Module (CBAM) [28]. These attention modules are applied at the same place, i.e. after the second convolution layer and the split-based inception block on the UTKFace dataset, modified with an artificial mask utilized from [2], to perform a fair comparison. The validation accuracy of the proposed architecture with RAM is higher than the proposed architecture with BAM, SE, or CBAM, which differ by 0.48%, 0.39%, and 0.12%, respectively.



(a)

(b)

Fig. 6. The correct detection results of the GenderMask-CPU detector for masked (a) and non-masked (b) faces.

Table 5. Comparisons of Different Attention Modules applied on the Proposed Architecture on UTKFace dataset, modified with an artificial mask utilized from [2].

Attention Modules	**Number of Parameters**	**GFLOPs**	**Validation Accuracy**	**Gender Recognition (FPS)**	**Gender Recognition integrated with Face Detection (FPS)**
BAM [14]	449,736	0.0348	90.69	238.19	125.40
SE [8]	440,946	0.0284	90.78	307.44	145.18
CBAM [28]	441,084	0.0286	91.05	273.84	135,84
RAM (ours)	**441,460**	**0.0293**	**91.17**	272.80	135.02

6 Conclusion

An efficient CPU-based human face detector with gender identification called GenderMask-CPU is proposed and offers a lightweight architecture with a split-based inception block and regulated attention module. This lightweight architecture assigns a few convolution operations that make the architecture only generates 441,460 parameters. This work offered a split-based inception block to efficiently extract features at various sizes by partially applying different convolution kernel sizes and levels. The regulated attention module is also proposed to improve the quality of the feature map. This architecture acquires competitive performance compared to other lightweight and deep CNN architectures on the UTKFace and Labeled Faces in the Wild (LFW) datasets, modified with an artificial mask utilized from [2]. Accordingly, when operating on a CPU device in real-time, GenderMask-CPU is capable of running at 135 frames per second while identifying the gender of masked faces. This detector outperforms other lightweight and deep competitors' architecture. In a forthcoming study, other mechanisms, such as Transformer, can be conducted to improve the identification accuracy. The augmentation strategy can also be explored to improve the dataset varieties that can increase the performance of masked face gender recognition.

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. Alhalabi, M., Hussein, N., Khan, E., Habash, O., Yousaf, J., Ghazal, M.: Sustainable smart advertisement display using deep age and gender recognition. In: 2021 International Conference on Decision Aid Sciences and Application (DASA). pp. 33–37. IEEE (2021)
2. Anwar, A., Raychowdhury, A.: Masked face recognition for secure authentication. arXiv preprint arXiv:2008.11104 (2020)
3. Bandung, Y., Hendra, Y.F., Subekti, L.B.: 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). pp. 1–6. IEEE (2015)
4. Greco, A., Saggese, A., Vento, M.: Digital signage by real-time gender recognition from face images. In: 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT. pp. 309–313. IEEE (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
7. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)
8. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(8), 2011–2023 (2019)

9. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(8), 2011–2023 (2020). https://doi.org/10.1109/TPAMI.2019.2913372
10. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: 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)
11. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
12. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
13. Mishima, K., Sakurada, T., Hagiwara, Y.: 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). pp. 573–575. IEEE (2017)
14. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514 (2018)
15. Priadana, A., Maarif, M.R., Habibi, M.: Gender prediction for instagram user profiling using deep learning. In: 2020 International Conference on Decision Aid Sciences and Application (DASA). pp. 432–436. IEEE (2020)
16. Priadana, A., Putro, M.D., Jeong, C., Jo, K.H.: A fast real-time face gender detector on cpu using superficial network with attention modules. In: 2022 International Workshop on Intelligent Systems (IWIS). pp. 1–6 (2022). https://doi.org/10.1109/IWIS56333.2022.9920714
17. Priadana, A., Putro, M.D., Jo, K.H.: An efficient face gender detector on a cpu with multi-perspective convolution. In: 2022 13th Asian Control Conference (ASCC). pp. 453–458 (2022). https://doi.org/10.23919/ASCC56756.2022.9828048
18. Priadana, A., Putro, M.D., Vo, X.T., Jo, K.H.: An efficient face-based age group detector on a cpu using two perspective convolution with attention modules. In: 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR). pp. 1–6. IEEE (2022)
19. Putro, M.D., Nguyen, D.L., Jo, K.H.: Lightweight convolutional neural network for real-time face detector on cpu supporting interaction of service robot. In: 2020 13th International Conference on Human System Interaction (HSI). pp. 94–99. IEEE (2020)
20. Putro, M.D., Priadana, A., Nguyen, D.L., Jo, K.H.: 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). pp. 171–178 (2022). https://doi.org/10.1109/AIM52237.2022.9863289
21. Ranjan, R., Patel, V.M., Chellappa, R.: 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 **41**(1), 121–135 (2017)
22. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018). https://doi.org/10.1109/CVPR.2018.00474
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4510–4520. IEEE (2018)

24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826. IEEE (2016)
28. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
29. Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D.: Epsanet: An efficient pyramid squeeze attention block on convolutional neural network. In: Proceedings of the Asian Conference on Computer Vision. pp. 1161–1177 (2022)
30. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5810–5818 (2017)