**Journal of Industrial Information Technology and Application** *ISSN(Online): 2586-0852*
Vol.6 No.4

# Development of Face and Landmark Detection using EfficientNetV2 and BiFPN

*Hyunduk Kim *, Sang-Heon Lee, Myoung-Kyu Sohn*

Division of Automotive Technology, DGIST, Daegu, Republic of Korea

**Abstract**. In this paper, we develop improved face and landmark detection algorithm using EfficientNetV2 as backbone and BiFPN as multi-scale feature extractor. EfficientNetV2 are a new family of convolutional networks that have faster training speed and better parameter efficiency than previous models. BiFPN is a new multi-scale feature extractor that allows easy and fast multi-scale fusion. To evaluate the performance of the proposed face and landmark detection algorithm, we trained and tested on WIDER FACE dataset using PyTorch framework in NVIDIA Titan RTX GPU. In the experiments, we show the experimental results for comparing the detection accuracy and efficiency of the proposed network to MobileNetV1 0.25 and Resnet50 networks. The experimental results show that the proposed algorithm is accurate and efficient than previous works.

**Keywords;** Face Detection; Face landmark detection; EfficientNetV2; BiFPN

## 1. Introduction

Accurate face and landmark detection are the essential preprocess of face analysis for many applications, such as facial identity and attribute recognition. In this paper, we develop improved face and landmark detection algorithm using EfficientNetV2 and BiFPN. In ICML 2021 conference, Tan and Le [1] introduced a new family of convolutional networks, EfficientNetV2 that have faster training speed and better parameter efficiency than previous models. They used new operator such as Fused-MB Conv in early layers and progressive learning, which adaptively adjusts regularization along with image size. They searched new model EfficientNetV2-S using Mnasnet [2] and scaled up EfficientNetV2-S to get EfficientNetV2-M/L. In CVPR 2020 conference,

* Corresponding author: hyunduk00@dgist.ac.kr
Received: Feb 9, 2022; Accepted: Feb 28, 2022; Published: Dec 30, 2022

Tan et al. [3] introduced a weighted bi-directional feature pyramid network (BiFPN), which allows easy and fast multi-scale feature fusion. The main ideas for BiFPN are efficient bidirectional cross-scale connections and weighted feature fusion. In this paper, we apply EfficientNetV2 as backbone network and BiFPN as multi-scale feature extractor for efficient and accurate face and landmark detection. Fig. 1 shows the overview of the proposed face and landmark detection approach.

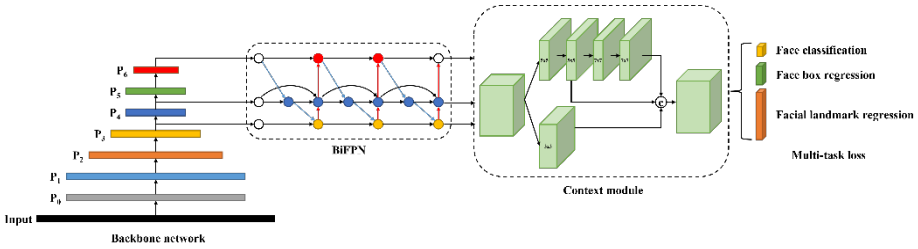


Fig 1. An overview of the proposed face and landmark detection approach.

## 2. Methodology

In the previous study, we used MobileNetV1 0.25 and Resnet50 as backbone, and FPN as multi-scale feature extractor similar to RetinaFace [4]. However, they cannot detect small or occluded face. Hence, we apply EfficientNetV2 as backbone and BiFPN as multi-scale feature for efficient and accurate face and landmark detection. EfficientNetV2 use Fused-MBConv in early stage 1-3 and MBConv in stage 4-6 to improve training speed with a small overhead on parameters. Fused-MBConv is proposed to better utilize mobile or server accelerators, which replaces the depthwise conv 3×3 and expansion conv 1×1 in MBConv with conv 3×3. Fig. 2(a) shows the structure of MBConv and Fused-MBConv. Generally, reducing the number of parameters means changing the model itself. We try to reduce the size of the model while keeping the overall model architecture. Starting from EfficientNetV2-S, we reduce the number of output channels and layers. Table I shows that the architecture for original EfficientNetV2-S and a lite version EfficientNetV2-SS. Moreover, we apply BiFPN for multi-scale feature fusion. As mentioned in [3], conventional FPN aggregates multi-scale features in a top-down manner, which is inherently limited by the one-way information flow. However, BiFPN has bidirectional (top-down and bottom-up) path for two-way information flow. Furthermore, BiFPN has an additional weight for each input to learn the importance of each input feature. In this paper, feature level 3, 4, and 6 are used in BiFPN and we stack three BiFPN layers for multi-scaling feature fusion. Fig. 2(b) shows the conventional top-down FPN and BiFPN multi-scale feature extractor. As shown in Fig. 1, we apply independent context modules on three feature pyramid levels

to increase the receptive field and enhance the rigid context modelling power. We use 3 × 3, 5 × 5, and 7 × 7 convolution layers and the outputs of these three layers are concatenated.

TABLE I. EFFICIENTNETV2-S/SS ARCHITECTURES

| Stage | Operator | EfficientNetV2-S | | | EfficientNetV2-SS | | |
|---|---|---|---|---|---|---|---|
| | | Stride | #Channels | #Layers | Stride | #Channels | #Layers |
| 0 | Conv 3x3 | 2 | 24 | 1 | 2 | 24 | 1 |
| 1 | Fused-MBConv1, k3x3 | 1 | 24 | 2 | 1 | 24 | 2 |
| 2 | Fused-MBConv4, k3x3 | 2 | 48 | 4 | 2 | 40 | 4 |
| 3 | Fused-MBConv4, k3x3 | 2 | 64 | 4 | 2 | 48 | 4 |
| 4 | MBConv4, k3x3, SE0.25 | 2 | 128 | 6 | 2 | 104 | 6 |
| 5 | MBConv6, k3x3, SE0.25 | 1 | 160 | 9 | 1 | 128 | 9 |
| 6 | MBConv6, k3x3, SE0.25 | 2 | 272 | 15 | 2 | 208 | 14 |
| 7 | Conv1x1 & Pooling & FC | - | 1280 | 1 | - | 1280 | - |

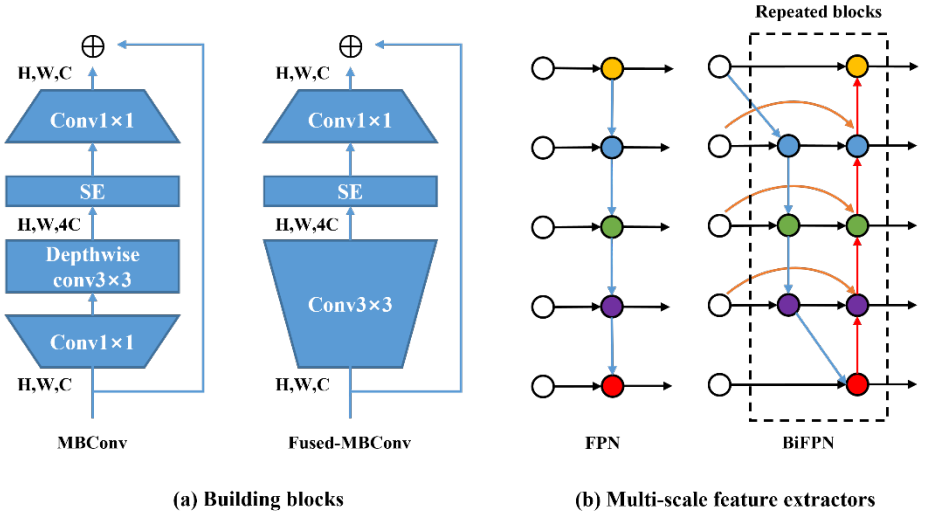


Fig 2. Structure of building blocks and multi-scale feature extractors.

## 3. Experiments

To evaluate the performance of the proposed face and landmark detection algorithm, we trained and tested on WIDER FACE dataset [5] using PyTorch framework in NVIDIA Titan RTX GPU. The WIDER FACE dataset comprises three levels of difficulty (Easy, Medium and Hard). All experiments are done on the NVIDIA Titan RTX using PyTorch. We train the proposed network using SGD optimizer (momentum at 0.9, weight decay at 0.0005). The learning rate starts from 0.0001, rising to 0.001 after 5 epochs, then divided by 10 at 190 and 220 epochs. The training process terminates at 250 epochs. We set the input image size to 640×480. For testing on WIDER FACE, we follow the standard protocol of the WIDER FACE dataset. Box voting is applied on the union set of predicted face boxes using an IoU threshold at 0.4.

We compare the detection accuracy and efficiency of the proposed network to MobileNetV1 0.25 and Resnet50 networks. As shown in Table II and III, EfficientNetV2-S/SS networks are more accurate than MobileNetV1 0.25. While these networks have less parameter than ResNet50, achieve similar accuracy. We can also observe that the conventional top-down FPN is inherently limited by the one-way information flow and thus has the lower accuracy than BiFPN. These results suggest that EfficientNet backbones and BiFPN are more efficient and accurate than previous approaches.

TABLE II. PERFORMANCE COMPARISON ON VARIOUS BACKBONES WITH FPN

| **+FPN** | | **MobileNetV1 0.25** | **ResNet50** | **EfficientNetV2 -S** | **EfficientNetV2 -SS** |
|---|---|---|---|---|---|
| **Total params** | | 426,608 | 27,293,600 | 21,871,848 | 12,617,252 |
| **Inference Time(sec)** | | 0.0200 | 0.1150 | 0.0918 | 0.0801 |
| **Accuracy** | **Easy** | 90.71 | 93.68 | **94.89** | 94.16 |
| | **Medium** | 88.17 | 92.49 | 93.51 | **93.53** |
| | **Hard** | 73.83 | 84.80 | **86.06** | 86.04 |

TABLE III. PERFORMANCE COMPARISON ON VARIOUS BACKBONES WITH BIFPN

| **+BiFPN** | | **MobileNetV1 0.25** | **ResNet50** | **EfficientNetV2 -S** | **EfficientNetV2 -SS** |
|---|---|---|---|---|---|
| **Total params** | | 821,872 | 33,593,248 | 22,267,112 | 13,012,516 |
| **Inference Time(sec)** | | 0.0208 | 0.1190 | 0.0955 | 0.0852 |
| **Accuracy** | **Easy** | 91.12 | 94.83 | **96.18** | 95.71 |
| | **Medium** | 89.54 | 93.40 | **95.32** | 95.06 |
| | **Hard** | 75.01 | 86.46 | **87.11** | 87.51 |

## 4. Conclusions

In this paper, we introduced improved face and landmark detection algorithm using EfficientNetV2 as backbone and BiFPN as multi-scale feature extractor. The experimental results showed that the proposed methods are accurate and efficient than previous works. In the future, we will apply vision transformer algorithm to obtain multi-task detection algorithm (landmark, head pose, and heart rate).

## Acknowledgment

This work was partly supported by the Technology development Program of MSS (S2860101) and the DGIST R&D Program of Ministry of Science and ICT (21-IT-02).

## References

[1] 2. M. Tan and Q. Le, “Efficientnetv2: Smaller models and faster training,” arXiv preprint arXiv:2104.00298, 2021.

[2] M. Tan, B. Chen, R. Pang, V. Vasudevan, and Q. V. Le, “Mnasnet: : Platform-aware neural architecture search for mobile,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2820-2828, 2019.

[3] 3. M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781-10790, 2020.

[4] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-stage dense face localisation in the wild,” arXiv preprint arXiv:1905.00641, 2019.

[5] 4. S. Yang, P. Luo, C. C. Loy, and X. Tang, “Wider face: A face detection benchmark,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5525-5533, 2016.