IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 13, 2021, accepted January 1, 2022, date of publication January 11, 2022, date of current version January 20, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3141791

Deep Learning-Based Facial Landmarks Localization Using Compound Scaling

SAVINA JASSICA COLACO AND DONG SEOG HAN, (Senior Member, IEEE)
School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding author: Dong Seog Han (dshan@knu.ac.kr)

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A03043144).

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Facial landmarks are crucial information needed in numerous facial analysis applications which can help to resolve difficult computer vision-related problems. The localization of landmarks, which involves facial keypoints such as eye centers, eyebrows, nose center, etc, offers necessary information for face analysis like expressions, emotions, health conditions, etc. The applications with requirement constraints such as the model size and computational load are often scaled up with better accuracy and efficiency. In this paper, we propose a deep learning-based approach for facial landmarks localization with compound dimension scaling. We modify the baseline network called EfficientNet with multi-scale fully connected layers to predict the facial landmarks on human faces which are mapped on the detected face in real-time. The proposed model with the compound scaling method gives a scalable model by uniformly scaling the width, depth and resolution dimensions. The model is evaluated with an adaptive wing loss function for both larger and smaller models. We also assessed the robustness of the model with various head poses and occlusion conditions. The proposed model which is trained with a large dataset can achieve 90% of accuracy for a larger model with a model size of 24.6 MB and approximately 88%∼89% of accuracy for smaller models. Hence, the smaller models can still achieve acceptable accuracy compared to the larger model with fewer parameters.

INDEX TERMS Adaptive wing loss, compound scaling, facial landmarks.

I. INTRODUCTION

The ability to recognize faces is an easy task for a human but a challenging one in the computer vision field. In particular, face recognition is less accurate compared to the identification methods such as fingerprint and iris scans. Face recognition can be defined as the method of finding and matching the faces in a given digital image or video against a database of faces. Facial landmarks localization or face alignment is one of the most researched areas in computer science that can improve the face recognition task. Facial landmarks localization detects prominent facial points like eyebrows, eye corners, nose, mouth corners, etc on the human faces of images [1]. Facial landmarks localization can help to identify positions of key points on the human's face where the information can be used in applications like driver assistance and monitoring systems. Since one of the factors resulting in fatal accidents listed by the U.S. Department of Transportation is driver attention [2]. To reduce such fatal accidents, we could monitor the driver's status information such as head pose variations, gaze movement or emotion using facial landmarks.

Due to the remarkable research done on deep learning, many research works show great accuracy on challenging computer tasks. The deep learning approaches have been popular in recent years and show better performance than the conventional approaches. The convolutional neural networks (CNNs) for computer vision tasks such as facial landmarks localization have made a lot of advancements over the past few years. CNN based facial landmarks localization gets the high-level features from the face and predicts all the facial keypoints. With CNN, the facial keypoints are detected

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

and localized simultaneously. Since there are tremendous contributions to computer vision tasks using CNN to achieve state-of-the-art performance, scaling CNN can also give better accuracy while keeping the model light and efficient. For example, the residual neural network (ResNet) [3] can be scaled down using network depth from ResNet-18 to ResNet-200, while MobileNets and WideResNet [4] can be scaled down using network width. Both network depth and width are important for CNN, but it is still an open problem as to how effectively CNN can be scaled for better efficiency and accuracy. Another popular method is to scale up the models by image resolution. It is also possible to scale two or three dimensions randomly requiring manual tuning and still produce sub-optimal accuracy and efficiency. The CNN needs more layers and channels for bigger input image size to increase receptive fields and to capture more fine-grained patterns on bigger images. Hence, it increases the complexity due to the overhead of more FLOPs for better accuracy. As applications specifically, driver monitoring assistance systems have constraints such as model size and computing requirements, building a scalable model can help to solve these limitations. Compound scaling can be used to scale all dimensions of the network to achieve a small model size and fast processing speed for the applications. To achieve these goals for facial landmarks localization, we propose a scalable model using a baseline EfficientNet also with multi-scale fully connected layers to have less computational load and increased accuracy with a smaller number of parameters. The main contributions of the paper are as follows:

1) We present a scalable and lightweight model for facial landmarks detection using baseline EfficientNet which uses compound model scaling approach to scale width, depth and resolution dimensions.
2) A multi-scale fully connected layer is added to extend the single-scale feature map which better extracts the global features by focusing on different face regions.
3) For better adaption to different pixel intensities, we evaluate the model with adaptive wing loss, which decreases small errors on foreground pixels for better landmark localization, while tolerating small background pixels with improved convergence rate.
4) Experiments are conducted on commonly used face alignment benchmarks such as 300W and 300VW where the proposed model gives better results compared to the baseline model.

The following sections of this paper are organized as follows: Section 2 discusses the related work on facial landmarks localization using different algorithms. Section 3 presents the proposed model used for facial landmarks localization. Real-time results and discussion are given in Section 4. Finally, we conclude in Section 5.

II. RELATED WORKS

This section introduces related works briefly on face detection, face alignment, landmark localization and convolution design.

A. FACE DETECTION

Among different proposed face detection methods, multi-task cascaded convolutional networks (MTCNN) [5] and faster region-based convolutional neural networks (R-CNN) [6] are well-known. Many object detection methods popularly use faster R-CNN where it generates bounding boxes based on predefined anchors with object classification. Subsequently, it crops the feature maps with object detection and refines the bounding box proposals for better results. Faster R-CNN with ResNet as a backbone is used for face detection.

B. FACE ALIGNMENT

The face alignment methods can be categorized as hand-crafted-feature-based and deep-learning-based methods. In the hand-crafted-feature-based methods, the tree structure part model (TSPM) used a deformable part-based model for face shape modeling with a mixture of other tree models for landmark localization, pose estimation and detection in parallel. For capturing the face appearance, cascade regression-based methods with scale-invariant feature transform (SIFT) features were used but the methods were incapable to find an unrestricted face with extreme poses. The statistical methods such as the constrained local model (CLM) and active appearance model (AAM) maximize the confidence of keypoints in an image. A real-time face aligning with facial landmarks using an ensemble of regression trees in [7]. The drawbacks of conventional approaches are expensive computation, high complexity of the model and less robustness.

C. FACIAL LANDMARKS LOCALIZATION

Deep learning-based approaches have been adopted since they outperform the conventional approaches. We briefly introduce the works in landmarks localization. A multi-task learning network [8], called tasks-constrained deep convolutional network (TCDCN), learns pose attributes and landmarks locations jointly. TCDCN is difficult to train because of its multi-task approach. Trigeorgis *et al.* [9] proposed a model for facial alignment with an end-to-end recurrent convolution from coarse to fine, where the model is termed as mnemonic descent method (MDM). Lv *et al.* [10] proposed an architecture with the two-stage re-initialization (TSR) scheme using deep regression, which boosts detection accuracy by dividing the whole face into several parts. Yang *et al.* [11] proposed a network for landmarks detection which is assisted by head pose angles including yaw, pitch and roll. Jourabloo and Liu [12] proposed a pose-invariant face alignment (PIFA) model which estimates a 3D to 2D projection matrix using deep cascade regressors which later extended with the convolutional neural network [13]. Zhu *et al.* [14] proposed a model with the face depth in Z-buffer and later fits 2D images in a 3D model. Kumar and Chellappa [15] designed a convolution neural network with a single dendritic, named pose conditioned dendritic convolution neural network (PCD-CNN), for the improvement

of detection accuracy by associating classification network with modular and second classification networks. Honari *et al*. [16] designed sequential multitasking (SeqMT) network equivariant landmark transformation (ELT) loss term. Dong *et al*. [17] created the style-aggregated network (SAN) for robust face landmark detection with the intrinsic variance of image styles. Wu *et al*. [18] proposed a face alignment algorithm considering the boundary information as a geometric structure attribute for human faces. The landmarks are derived from boundary information to avoid ambiguities. Fan and Zhou [19] use multiple CNNs to increase robustness and improve accuracy for coarse-to-fine prediction. Weng *et al*. [20] proposed a feature set matching approach for partial face matching by explicitly constrain the affine matrix. Ranjan *et al*. [21] proposed a novel architecture called HyperFace by fusing the intermediate layer of the CNN with post-processing methods namely iterative region proposals and landmarks-based non-maximum suppression to improve the overall performance. Xiao *et al*. [22] developed a recurrent dual refinement model (RDR) to focus on large-pose facial landmark detection with an end-to-end framework. Lai *et al*. [23] proposed a novel recurrent network to refine the estimated coordinates of facial landmarks iteratively. Junfeng and Haifeng [24] presented a global exemplar stacked auto-encoder network (GECSAN) face shape initialization and local information preserve stacked auto-encoder networks (LIPSAN) for shape refinement to achieve a robust face refinement. Xia *et al*. [25] proposed a CNN based head pose estimation with facial landmarks using heatmaps. Kim *et al*. [26] proposed an extended multi-task CNN (EMTCNN) model for facial landmarks detection in real-time. Zheng *et al*. [27] presented a model that aims at providing an efficient coarse-to-fine network by utilizing lightweight coordinate regression and heatmap regression. Zhang *et al*. [28] developed a structural hourglass network to predict the facial landmarks with corresponding heatmaps. The CNN based facial landmarks localization gets the high-level features from the face and predicts all the keypoints simultaneously.

D. CONVOLUTION DESIGN

In various deep learning, CNN is more accurate with the bigger network such as AlexNet [29], which won the 2012 ImageNet competition while GoogleNet [30] achieved the top-1 accuracy of 74.8% with 6.8 M parameters at 2014 ImageNet, the top-1 accuracy of 82.7% by SENet [31] with 145 M parameters in 2017 ImageNet challenge. Some of the ImageNet models work well across various transfer learning datasets and other computer vision problems. It is crucial to achieving high accuracy for many applications but there are constraints with hardware memory. Hence an accuracy gain needs good efficiency. A common way to trade accuracy for efficiency is to reduce the model size or handcraft efficient mobile-size CNN such as SqueezeNets [32], MobileNets [33] and ShuffleNets [34]. By tuning the network depth, width, kernel types and sizes, the CNN can achieve better efficiency [35]. It is difficult to achieve this tuning in the larger model which has expensive tuning costs but can be achieved by the model scaling approach.

III. PROPOSED MODEL

A. COMPOUND MODEL SCALING

A CNN layer i can be defined as a function: $Y_i = F_i(X_i) + b$, where F_i is the operator, b is the bias term, Y_i is the output feature map with C_i channel dimension, X_i is the input tensor. The input tensor has (H_i, W_i, C_i) shape, where H_i and W_i are the spatial dimensions and C_i is the channel dimension. A CNN is a sequence of convolutional layers with other layers such as activation function, pooling, etc. A CNN can be partitioned into multiple stages and each stage can have similar architecture. For example, ResNet consists of 5 stages and each stage has the same convolutional type except down-sampling at the first layer. Most of the simple convolution neural network designs focuses on finding the best layer architecture F_i while the model scaling expands either by scaling the width, network length or resolution without changing the operator F_i in baseline network. The optimal values of the width, depth and resolution depend on each other and the values under different resource constraints change. Hence conventional methods can scale CNNs in one of the dimensions. The most common type of scaling in CNNs is network depth scaling. The assumption is made that a deeper network can capture features that are more complex and generalize well on new data. However, they are difficult to train due to the vanishing gradient problem. For instance, ResNet-101 has similar accuracy with ResNet-1000 though ResNet-1000 has a greater number of layers. The scaling width of the network is also commonly used in smaller models. Fine-grained features can be captured with wider networks and they are easier to train. Nevertheless, shallow networks that are extremely wide have difficulty capturing high-level features [30]. Moreover, CNNs can capture more fine-grained features with higher resolution input images.

Gpipe [36] achieved the state-of-the-art accuracy for ImageNet with 480 × 480 resolution. However, the accuracy gain diminishes for very high resolutions [35]. The scaling of any one of the dimensions improves accuracy, but accuracy reduces for larger models. Hence, the compound model scaling simplifies the design problem by scaling width, network depth and resolution. The scaling of these dimensions should be done uniformly with a constant ratio for better accuracy and efficiency. The dimensions are scaled in a principled way,

$$\begin{aligned} \text{Depth: } d &= \alpha^{\phi}, \\ \text{Width: } &= w\,\beta^{\phi}, \\ \text{Resolution: } r &= \gamma^{\phi}, \\ \text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2, \\ \alpha &\geq 1,\ \beta \geq 1, \gamma \geq 1 \end{aligned} \tag{1}$$

where the dimensions of α, β and γ in (1) are decided by the grid search method, which these dimensions are taken as constant values. A user-specified coefficient, ϕ, control how

many of the resources are needed for model scaling. The α, β and γ reveal how the extra resources are assigned to network width, depth and resolution, respectively. The desired target model size or computational cost are achieved by getting the coefficients through the grid search. All layers are restricted to reduce the design space which allows the layers to be scaled uniformly with a constant ratio.

A small grid search allows us to decide the dimensions of α, β and γ in (1), which are constant values. A user-specified coefficient, ϕ, control how many of the resources are needed for model scaling. The α, β and γ reveal how the extra resources are assigned to network width, depth and resolution, respectively. The desired target model size or computational cost are achieved by getting the coefficients through grid search. All layers are restricted to reduce the design space which allows the layers to be scaled uniformly with a constant ratio.

B. PROPOSED FACIAL LANDMARKS LOCALIZATION

A neural search architecture develops a baseline EfficientNet with the AutoML mobile neural architecture search (MNAS) framework, which enhances both accuracy and efficiency [35]. The architecture developed by the neural search architecture utilizes the mobile inverted bottleneck convolution (MBConv), which are similar to MobileNetV2 [37] and MnasNet [38]. The EfficientNet model uses the concept of compound scaling to scale the network depth, width and image resolution without changing the architecture's predefined baseline network to improve the overall performance. Fig. 1 depicts a comparison between simple CNN and compound scaling. The spatial dimensions $H \times W$ of simple CNN are gradually minimized and channel dimension C is expanded over layers. Furthermore, a grid search is done for the compound scaling method by finding a correlation between scaling dimensions while keeping the resource constraints fixed for the baseline network. Table 1 provides the architecture configurations of the baseline EfficientNet model.

The proposed model uses compound model scaling which helps to maintain the resources constraints to achieve less computation load and more accuracy. The proposed facial landmarks model adopts a baseline EfficientNet model with an input size of 112×112 with grayscale. The EfficientNet uses the concept of compound model scaling to scale the width, depth and resolution uniformly. Furthermore, the EfficientNet optimizes both accuracy and FLOPS by leveraging a multi-objective neural architecture. The core building block in EfficientNet is MBConv, which takes two inputs, the first input is the data and the other input is block arguments. The data is output from the last layer and a block argument is a collection of attributes used in MBConv like input filters, output filters, expansion ratio, squeeze ratio, etc. In the MBConv block, the expansion ratio expands the layer and make them wide hence, the connected blocks are narrower and inner blocks are wider by increasing the number of channels. After expanding the layer, depthwise convolution applies a single

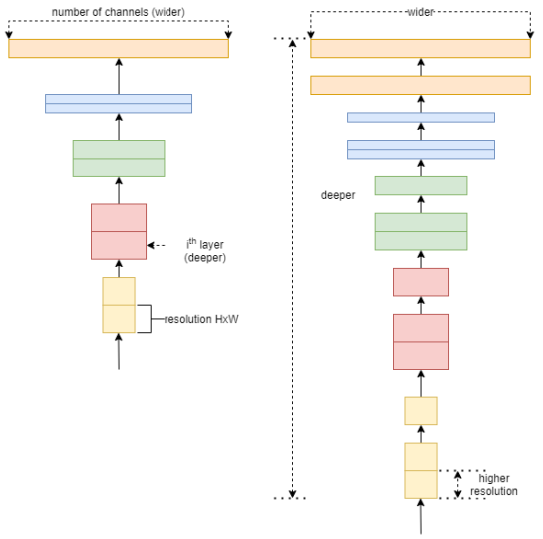


FIGURE 1. Model scaling. (Left) simple CNN with one of the dimensions of network width, depth or resolution is increased (Right) Compound scaling uniformly scales all three dimensions. [35].

filter per each input channel with a definite kernel size. Furthermore, the squeeze ratio is used to squeeze the number of channels.

Since human faces have strong global structures, like symmetry and spatial relationships among eyes, mouth, nose, etc., these global structures could help to localize landmarks more precisely. Instead of single-scale feature maps, we extend them into multiple-scale maps to enlarge the receptive field and better catch these global features on faces. Therefore, an MS-FC layer is added in the proposed architecture for precisely localizing landmarks in images. This added layer also helps to increase the accuracy of the baseline EfficientNet. We replace the convolutional layers in MS-FC with a compound model scaling method to make these layers scalable uniformly. In Table 2, the multi-scale fully connected layers are donated by S1, S2 and S3 and later concatenated. Fig. 2 shows the proposed architecture with EfficientNet and MS-FC layers. The activation function used in the model is Hswish instead of swish. Though nonlinearity improves accuracy by swish, it comes with a non-zero cost as sigmoid is expensive to compute on mobile devices. Hence, we use Hswish by replacing the expensive sigmoid with its piece-wise linear hard analog: $\text{ReLU6}(x+3)/6$. The rectified linear unit (ReLU) such as ReLU6 is used instead of a custom clipping constant. The hard version of the swish is

$$\text{Hswish}(x) = x\frac{\text{ReLU6}(x+3)}{6} \tag{2}$$

C. LOSS FUNCTION

The model is evaluated with the adaptive wing (Awing) loss [39] which was mainly used for heatmap regression as

$$\text{Awing}(y,\hat{y}) = \begin{cases} \omega \ln\left(1 + \left|\frac{y-\hat{y}}{\varepsilon}\right|^{\alpha-y}\right), & \text{if } |(y-\hat{y})| < \theta \\ A\,|(y-\hat{y})| - C, & \text{otherwise} \end{cases} \tag{3}$$

TABLE 1. Baseline EfficientNet model architecture configurations [35].

Stage	Name	Input size	Output channels	Layers
1	Conv3 × 3	224 × 224	32	1
2	MBConv1, k3 × 3	112 × 112	16	1
3	MBConv6, k3 × 3	112 × 112	24	2
4	MBConv6, k5 × 5	56 × 56	40	2
5	MBConv6, k3 × 3	28 × 28	80	3
6	MBConv6, k5 × 5	14 × 14	112	3
7	MBConv6, k5 × 5	14 × 14	192	4
8	MBConv6, k3 × 3	7 × 7	320	1
9	Conv1 × 1, Pooling, FC	7 × 7	1280	1

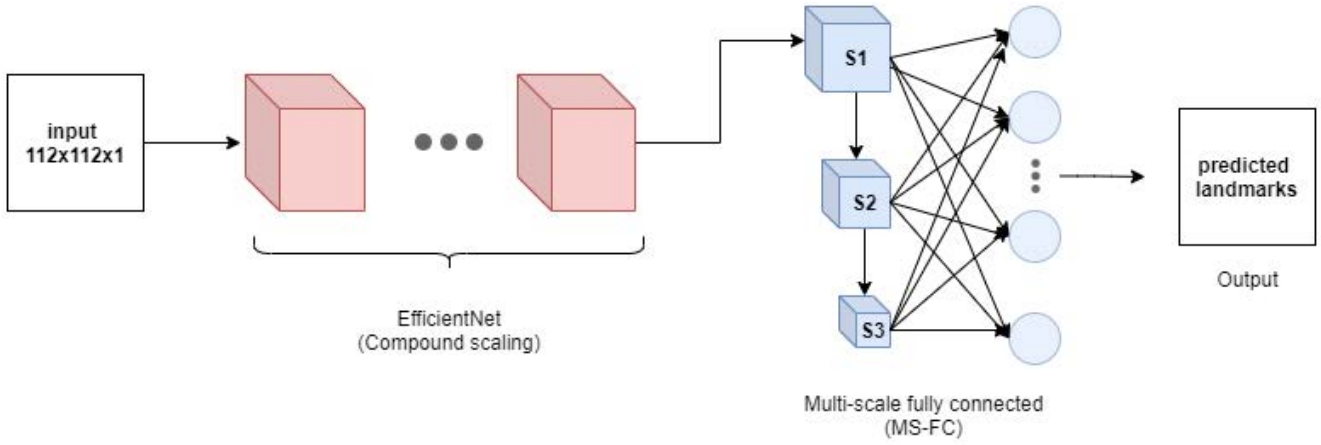


FIGURE 2. Proposed model design for facial landmarks detection with modified baseline network and MS-FC layers.

TABLE 2. Proposed model architecture configurations for facial landmarks prediction.

Stage	Name	Input size	Channels	Layers	Stride
1	Conv3 × 3	112 × 112	32	1	2
2	MBConv1, k3 × 3	55 × 55	16	1	1
3	MBConv6, k3 × 3	55 × 55	24	2	2
4	MBConv6, k5 × 5	28 × 28	40	5	2
5	MBConv6, k3 × 3	14 × 14	80	3	2
6	MBConv6, k5 × 5	7 × 7	112	3	1
7	MBConv6, k5 × 5	7 × 7	192	4	2
8	MBConv6, k3 × 3	4 × 4	320	1	1
(S1)	Conv	4 × 4	16	1	1
(S2)	Conv	4 × 4	32	1	1
(S3)	Conv	4 × 4	128	1	1
S1,S2,S3	Full Connection	-	136	1	-

where y and $\hat{y}$ are ground truth and predicted pixels values, respectively. Since the wing loss [40] use ω as a threshold, the adaptive wing loss introduces θ, a new variable threshold which is used to switch between linear and non-linear part. The values for ω, θ, ε and α are all positive. $A = \omega[1/\{1 + (\theta/\varepsilon)^{(\alpha-y)}\}](\alpha - y)\{(\theta/\varepsilon)^{(\alpha-y-1)}\}(1/\varepsilon)$ and $C = [\theta A - \omega \ln\{1+(\theta/\varepsilon)^{(\alpha-y)}\}]$ make the loss function smooth and continuous at $|y - \hat{y}| = \theta$. The error is considered small and strong influence is needed when $|y - \hat{y}| < \theta$. An exponential term $\alpha - y$ is used to make the loss function shape adapt with y and the loss function to be smooth at point zero. Similar settings from [39] are used for the experiment such as $\theta = 14$, $\theta = 0.5$, $\varepsilon = 1$ and $\alpha = 2.1$.

IV. RESULTS AND DISCUSSION

A. DATASET

The experiment is conducted on widely used datasets, 300W and 300VW [41]–[43] to predict facial landmarks. The 300W dataset [44] annotates five existing datasets such as XM2VTS, AFW, HELEN, LFPW and IBUG with 68 landmarks. 300VW is mainly designed as a benchmark for videos, containing 50 training videos and 64 testing videos. In this paper, a few samples of images from the 300VW dataset are combined with the 300W dataset. The few samples are taken considering the different scenarios in the 300VW dataset. The complete dataset used for the experiment consists of 112,111 images in total.

B. IMPLEMENTATION DETAILS

The experiments for both baseline and proposed networks are carried out with 112 × 112 resized grayscale images. Both networks are trained on Nvidia GeForce GTX 980Ti GPU. The Keras framework is used for the models with a batch size of 100 and tested with a different number of epochs. The Adam optimization method [45] is employed for network training, which is a useful stochastic optimization that only

FIGURE 3. Face detection using ResNet and SSD.

requires first-order gradients with less memory. It combines the advantages of two popular methods: AdaGrad [46], which is well with sparse gradients, and RMSProp [47], which works well in online and non-stationary settings. Adam is used instead of stochastic gradient descent to update network weights iterative based on training data. The learning rate of 10^{-3} is used with the Adam optimizer and reduced after every 20 epochs. The data are split with an 80% to 20% ratio with 89,688 images are used as training data and 22,423 images are used as testing data. The experiment is also conducted with a 300W dataset which is divided into training data of 3,148 images and test data of 689 images. The test images are further divided into two subsets such as a common subset of 554 images from LFPW and HELEN and a challenging subset of 135 images from IBUG. A full testing set is formed by combining both common and challenging subset

C. FACE DETECTION

ResNet and single-shot detector (SSD) [48] are employed to detect the faces in images or video frames. The single-shot detector is a one-stage detection algorithm that does not need an initial object proposals generation step. Hence, it is faster and more efficient than two-stage approaches such as Faster R-CNN [6]. The detected faces are mapped with facial landmarks predicted with the proposed model. The red box indicates the bounding box with detected faces in Fig. 3.

D. FACIAL LANDMARKS LOCALIZATION

1) EVALUATION OF COMBINED DATASET

The baseline EfficientNet was trained for a different number of epochs for instance 300, 500 and 900 to predict the landmarks shown in Fig. 4. For width = 1.0 and depth = 1.0, the model maintained around 86% accuracy with no dropout. The training and testing data converge after a certain number of epochs and generalizes well. However, for 900 epochs, the test data does not seem to converge well. Fig. 5 shows the real-time detection with different epochs. The real-time detection shows better landmark detection around the mouth region but not in the case for the model trained with 900 epochs. There is a lot of disruption with the landmarks with extreme head poses. One of the reasons could be not having enough images for different cases such as extreme head poses or illumination conditions. The model could overfit the data and hence there shows no improvement in the accuracy for the higher number of epochs. The number of epochs is set high as possible and the best values are saved based on loss values.

The scaling of CNN can give better accuracy while keeping the model light and efficient. For example, ResNet is scaled using the network depth dimension. Though ResNet improves the training, the gain accuracy of the accuracy tends to decrease with the deeper network. With applications having model size constraints, scaling only the width dimension could be difficult to capture high-level features which can result in quicker saturation of accuracy. For the evaluation of the proposed model, the baseline EfficientNet and the proposed model were tested with a width value of 0.5 and a depth value of 1.1 in Table 3. The baseline EfficientNet has the same number of layers excluding the MS-FC layers. Both the models were trained for 150 epochs. Even with a 0.2 million parameters increase due to MS-FC layers, the model with a width value of 0.5 improved the accuracy to 90% from 87%. The smaller model is still able to give better localization of landmarks than the larger model. Although the extension of feature maps made the proposed model deeper, the accuracy gain did not decrease rather increased it while maintaining the other dimensions.

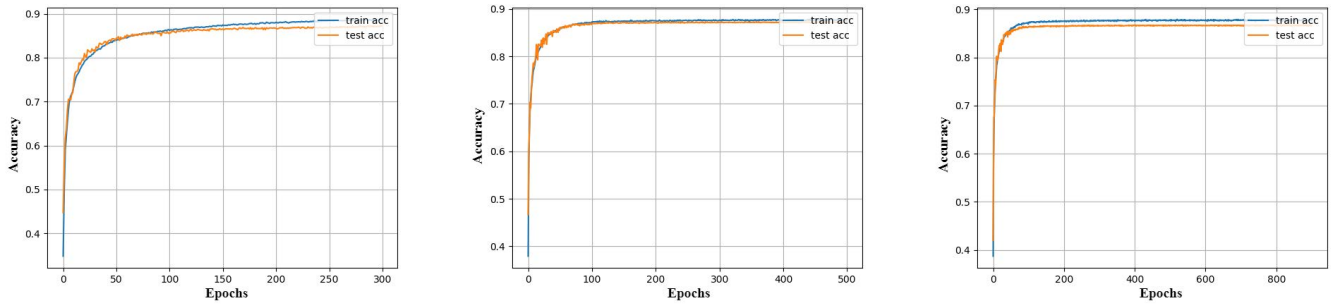
TABLE 3. Comparison with baseline EfficientNet and proposed model.

Model	Width	Depth	Parameters	Accuracy
Base EfficientNet	0.5	1.1	1.8M	87%
Proposed model	0.5	1.1	2M	90%

In Table 4, the proposed model with a width value of 0.5 and 0.25 was initially tested with different values of depth. As shown in Figs. 6 and 9, the smaller models gave better accuracy than the larger values. Figs. 8 and 10 shows real-time detection with a smaller width. The smaller maintain the accuracy of around 90% with a slight accuracy drop when a higher number of depth values. When the model depth values which in turn deepens the network but after a certain threshold the accuracy diminishes. Hence a uniform scaling should be done using compound model scaling with width, depth and resolution dimensions. Fig. 7 depicts the accuracy of width values of 0.5 and 0.25 with different depth values.

2) EVALUATION OF 300W DATASET

For the 300W dataset, we measure accuracy in terms of normalized mean error (NME), as done in previous works [8], [9], [12], where the normalized errors are averaged over all annotated landmarks. The proposed method is reported with an inter-pupil and interocular normalizing factor, which is normalized by the eye center and outer eye corners distance. The proposed model has modified baseline EfficientNet with MS-FC layers. The model is trained with a 300W dataset with three subsets such as common, challenging and full set. As shown in Tables 5 and 6, the proposed model is evaluated with different values of width and depth against normalized mean error as an accuracy measure. With the comparison of other facial landmarks localization methods, the proposed model has better results with a full set which is a combination of common and challenging subsets. But for the two cases i.e. width = 1.0, depth = 1.0 and width = 0.25, depth = 2.2, the NME error is higher for common subset compared to full set.



(a) 300 epochs (b) 500 epochs (c) 900 epochs

FIGURE 4. Model accuracy of base efficientNet model at different number of epochs.

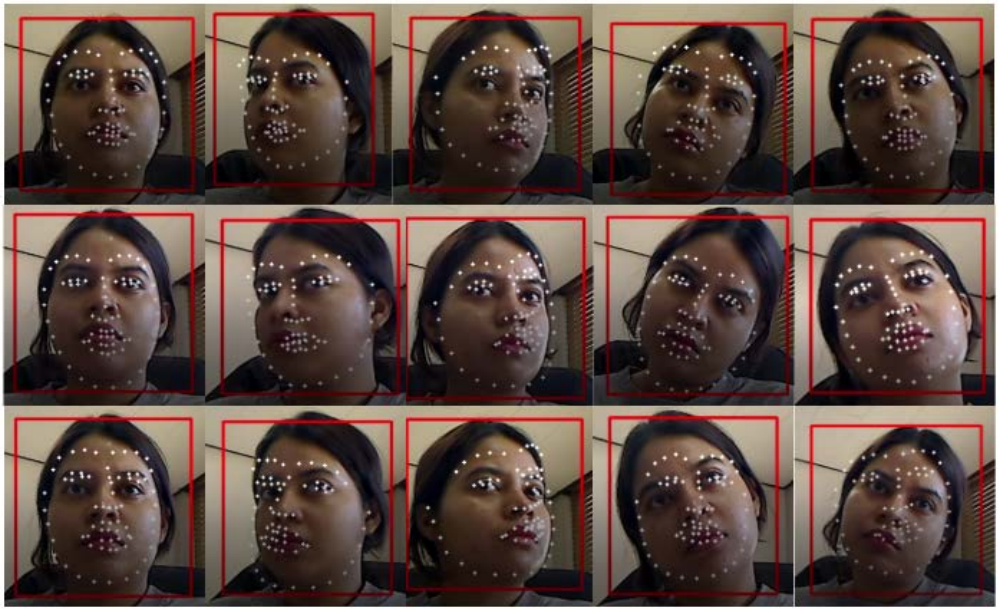


FIGURE 5. Real-time detection with top row at 300 epochs, middle row at 500 epochs and last row at 900 epochs.

TABLE 4. Comparison with baseline EfficientNet and proposed model for smaller width.

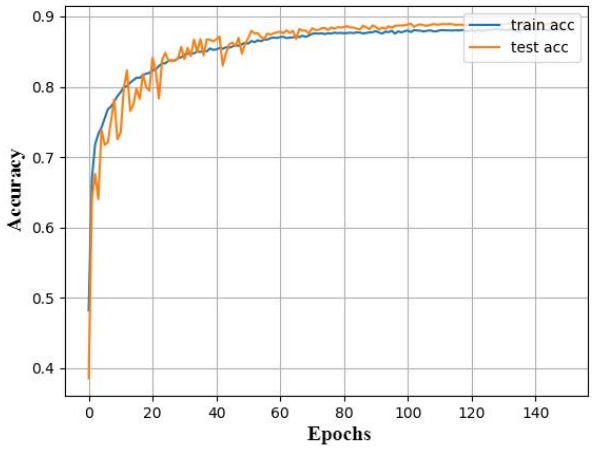
Model	Width	Depth	Parameters	Accuracy
Base EfficientNet	0.5	1.1	1.8M	87%
Proposed model	0.25	1.2	0.6M	88.9%
		2.2	1.1M	89.3%
		2.6	1.2M	88.4%

One of the reasons is due to the scaling of width and depth dimensions which needs to be uniformly managed with a constant ratio. Moreover, localization error could reach a certain threshold with different combinations of dimension values. The model with only a challenging subset has a higher error due to a small dataset of 135 images with an extreme variation. The challenging subset suffers from variations caused by head pose, facial expression and lower resolution. The proposed model gives an adequate NME value for interocular distance compared to the interpupil distance which could align facial landmarks around the eye corners rather than eye centres. The eye information could be vital information for estimating gaze or expression for the applications such as driver monitoring systems, emotion recognition etc.

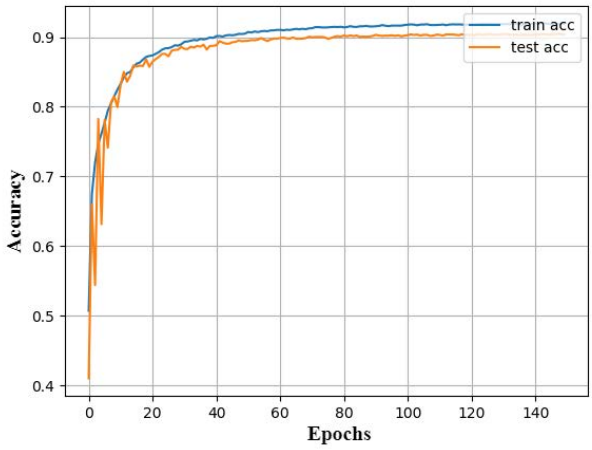
Although it is critical to balance all the dimensions of CNN, it could still achieve the balance between those dimensions by scaling each of them with a constant ratio [3], [32], [33], [36]. CNN can be scaled in a principle way with network width, depth or resolution dimensions [35].

3) PARAMETER AND MODEL SIZE

As shown in Table 7, the proposed EfficientNet for facial landmarks localization gives a small number of parameters compared to other CNN models. By expanding the CNN models, the number of parameters would also increase which could make the model deeper with diminishing accuracy gain [32]. The model trained with a smaller width is still able to achieve improved accuracy compared to the baseline model alone. The width, depth and resolution dimensions are uniformly scaled, giving higher accuracy with considerably fewer parameters. Hence model can be scaled to a smaller size and still achieve higher accuracy than the baseline and larger model. Model size in MB and frames per second (FPS) compared to other approaches could be shown in Table 8. Our proposed model gives 24.6 MB size at 30



(a) Baseline model



(b) Proposed model

FIGURE 6. Model accuracy of baseline EfficientNet and proposed model width = 0.5.

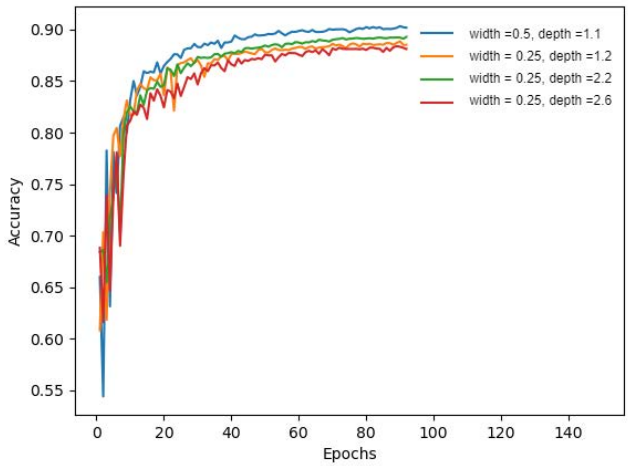


FIGURE 7. Accuracy for width values of 0.5 and 0.25 with different depth values.

FPS for the width and depth values of 0.5 and 1.1, respectively. This could be further reduced with width, depth or resolution values for the model which could be highly useful in mobile devices. The model size of SDM [59] shows a smaller model size but our proposed model has a smaller

TABLE 5. NME comparison with interpupil distance on 300W.

Method	Fullset
RCPR [49]	8.35
CFAN [54]	7.69
ESR [56]	7.58
SDM [59]	7.50
LBF [57]	6.32
CFSS [60]	5.76
3DDFA [53]	7.01
SHN-GCN [28]	**4.35**
Proposed (width=1.0, depth=1.0)	7.89
Proposed (width=0.5, depth=1.1)	7.87
Proposed (width=0.25, depth=1.2)	7.85
Proposed (width=0.25, depth=2.2)	8.15
Proposed (width=0.25, depth=2.6)	8.05

TABLE 6. NME comparison with interocular distance on 300W common subset, challenging subset and fullset.

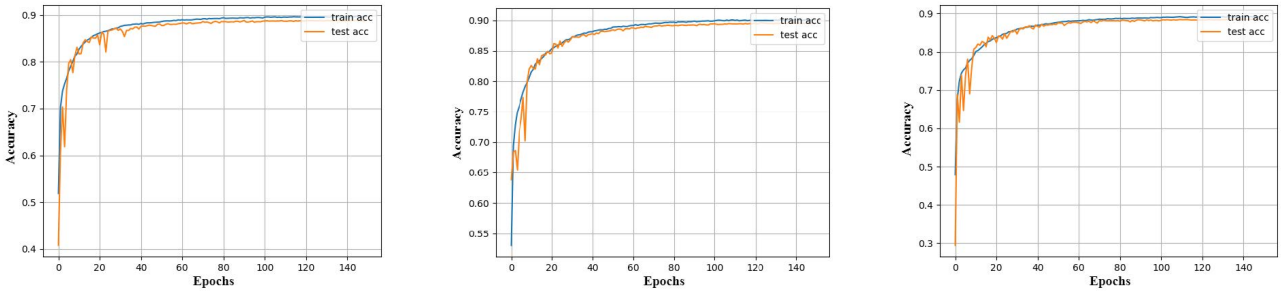
Method	Common	Challenging	Fullset
ESR [56]	5.28	17.00	7.58
RCPR [49]	6.18	17.26	8.35
SDM [59]	5.57	15.40	7.50
LBF [57]	4.95	11.98	6.32
TSR [12]	4.36	7.56	4.99
CFSS [60]	4.73	9.98	5.76
DSRN [50]	4.12	9.68	5.21
PCD-CNN [55]	3.67	7.62	4.44
RCFA [58]	4.03	9.85	5.32
RAR [51]	4.12	**8.35**	4.94
3DDFA [53]	6.15	10.59	7.01
PIFA-CNN [13]	5.43	9.88	6.30
Proposed (width=1.0, depth=1.0)	4.13	13.74	4.11
Proposed (width=0.5, depth=1.1)	4.09	13.88	4.10
Proposed (width=0.25, depth=1.2)	**4.05**	13.80	**4.07**
Proposed (width=0.25, depth=2.2)	4.24	13.71	4.15
Proposed (width=0.25, depth=2.6)	4.19	13.29	4.32

NME value in all subsets compared to SDM. The efficiency performance of TCDCN [8] is higher than our proposed model which processes 58 images per second but it only localizes 5-point landmarks rather than 68-point landmarks. Furthermore, CFAN [54] processes 40 images per second with 68-point landmarks prediction with $0.2 \sim 0.5$ NME error (interpupil distance) difference compared to our proposed model.

E. OCCLUSION

The robustness of the facial landmarks localization is evaluated with various head poses and occlusion. The landmarks are approximately localized with extreme head poses. The landmarks are detected well around the eye and nose area when occluded with only glasses. But tend to lose its shape when occluded with both glasses and a face mask. The eye region information such as the eye landmarks can be vital in the application such as driver assistance systems where we need to find driver' status information for stance gaze movement, mirror checking or lane checking. It can be also used to find the driver's physical states such as drowsiness, alertness, etc. The successful detection of faces could give a better detection of landmarks hence the face detector could be replaced with better detectors. Fig. 11 shows the detection with occlusion for the model accuracy at 90%.

FIGURE 8. Real-time detection with width = 0.5. (top) Baseline model (bottom) Proposed model.



(a) Depth value=1.2 (b) Depth value=2.2 (c) Depth value=2.6

FIGURE 9. Model accuracy of proposed efficientNet model with width = 0.25 and different depth values.

FIGURE 10. Real-time detection with width = 0.25. Depth values with 1.2, 2.2, and 2.6 at respective rows.

FIGURE 11. Facial landmarks localization with occulsion test with reading glasses and face mask.

V. CONCLUSION

This paper proposed a model to localize facial landmarks using the compound model scaling and multi-scale fully connected layers. The 68 predicted facial landmarks are mapped onto the human face shape. From the baseline EfficientNet model, the proposed model was able to better predict approximate facial key points on the input face in real-time. The proposed model for a smaller model achieves improved accuracy compared to the baseline EfficientNet. The model is also observed with a loss function called adaptive wing loss.

TABLE 7. Comparison with number of parameters with different CNN models used for facial landmarks localization.

Model	Input size	Parameters
VGG-19 [52]	224×224	143,667,240
HyperFace [21]	227×227	29,677,932
ResNet-18 [3]	224×224	11,689,512
Augmented EMTCNN [26]	–	6,083,186
Baseline EfficientNet (width=1.0,depth=1.0) [35]	112×112	4,320,194
Baseline EfficientNet (width=0.5,depth=1.1) [35]	112×112	1,887,648
Proposed model (width=1.0,depth=1.0)	112×112	4,368,514
Proposed model (width=0.5,depth=1.1)	112×112	2,011,776
Proposed model (width=0.25,depth=1.2)	112×112	682,012
Proposed model (width=0.25,depth=2.2)	112×112	1,105,108
Proposed model (width=0.25,depth=2.6)	112×112	1,243,780

TABLE 8. Comparison with model size and frames per second (FPS) with different state-of-the-art approaches.

Method	Points	Model Size	FPS
TCDCN [8]	5	-	58
HyperFace [21]	21	-	5
SDM [59]	68	10.1	-
SAN [17]	68	270.5+528	-
3DDFA [53]	68	-	13
CFAN [54]	68	-	40
LAB [18]	68	50.7	16
Proposed model (width=0.5,depth=1.1)	68	24.6	30

Although with improved accuracy, few facial features are not localized well due to insufficient data dealing with extreme conditions. The proposed model does not cope well with a severe variation of head poses and complete occlusion. In real-time, the facial landmarks localization could fail to detect the key points due to the failure of the face detector. In the future, we will focus on structuring a better model bearing in mind various imaging conditions. With an improved model, we can study important information for instance the head orientation of the user, which can be vital for the applications such as driver monitoring systems. For more diversity in data, data augmentation can be used to have data on various head poses, illumination and occlusions conditions, which can immensely improve the models. Boundary aware methods can help in the accurate alignment of landmarks. For the robust prediction of landmarks, a custom loss function can be encouraged. A better face detector can be utilized or combined with facial landmarks localization to improve the overall performance.

REFERENCES

[1] J.-K. Park and D.-J. Kang, "Unified convolutional neural network for direct facial keypoints detection," *Vis. Comput.*, vol. 35, pp. 1615–1626, May 2018.
[2] National Highway Traffic Safety Administration, *Traffic Safety Facts 2017: A Compilation of Motor Vehicle Crash Data*, DOT document vol. 812806, 2019.
[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[4] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
[5] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[6] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 650–657.
[7] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.
[8] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. ECCV*, 2014, pp. 94–108.
[9] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4177–4187.
[10] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3691–3700.
[11] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson, "Face alignment assisted by head pose estimation," 2015, *arXiv:1507.03148*.
[12] A. Jourabloo and X. Liu, "Pose-invariant 3D face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3694–3702.
[13] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3219–3228.
[14] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.
[15] A. Kumar and R. Chellappa, "Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 430–439.
[16] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Improving landmark localization with semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1546–1555.
[17] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 379–388.
[18] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2129–2138.
[19] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image Vis. Comput.*, vol. 47, pp. 27–35, Mar. 2016.
[20] R. Weng, J. Lu, and Y.-P. Tan, "Robust point set matching for partial face recognition," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1163–1176, Mar. 2016.
[21] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.
[22] S. Xiao, J. Feng, L. Liu, X. Nie, W. Wang, S. Yan, and A. Kassim, "Recurrent 3D-2D dual learning for large-pose facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1642–1651.
[23] H. Lai, S. Xiao, Y. Pan, Z. Cui, J. Feng, and C. Xu, "Deep recurrent regression for facial landmark detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1144–1157, May 2016.
[24] J. Zhang and H. Hu, "Exemplar-based cascaded stacked auto-encoder networks for robust face alignment," *Comput. Vis. Image Understand.*, vol. 171, pp. 95–103, Jun. 2018.
[25] J. Xia, L. Cao, G. Zhang, and J. Liao, "Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks," *IEEE Access*, vol. 7, pp. 48470–48483, 2019.

[26] H.-W. Kim, H.-J. Kim, S. Rho, and E. Hwang, "Augmented EMTCNN: A fast and accurate facial landmark detection network," *Appl. Sci.*, vol. 10, no. 7, p. 2253, Mar. 2020.
[27] S. Zheng, X. Bai, L. Ye, and Z. Fang, "HafaNet: An efficient coarse-to-fine facial landmark detection network," *IEEE Access*, vol. 8, pp. 123037–123043, 2020.
[28] J. Zhang, H. Hu, and S. Feng, "Robust facial landmark detection via heatmap-offset regression," *IEEE Trans. Image Process.*, vol. 29, pp. 5050–5064, 2020.
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.
[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[31] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[32] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.
[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[34] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[35] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[36] Y. Huang, Y. Cheng, A. Bapna, and O. Firat, "Gpipe: Efficient training of giant neural networks using pipeline parallelism," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 103–112.
[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[38] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2815–2823.
[39] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6970–6980.
[40] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2235–2245.
[41] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "Offline deformable face tracking in arbitrary videos," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 954–962.
[42] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-Wild challenge: Benchmark and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 50–58.
[43] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3659–3667.
[44] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-Wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 397–403.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[46] J. C. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 2121–2159, 2011.
[47] T. Tieleman and G. Hinton, "Lecture 6.5-RMSprop, coursera: Neural networks for machine learning," Univ. Toronto, Tech. Rep., 2012, pp. 26–31.
[48] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.
[49] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.
[50] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," in *Proc. ECCV*, 2018, pp. 585–601.
[51] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Proc. ECCV*, 2016, pp. 57–72.
[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[53] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1233–1245, Mar. 2016.
[54] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment," in *Proc. ECCV*, 2014, pp. 1–16.
[55] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, and H. Huang, "Direct shape regression networks for end-to-end face alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5040–5049.
[56] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 177–190, 2014.
[57] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1685–1692.
[58] W. Wang, S. Tulyakov, and N. Sebe, "Recurrent convolutional face alignment," in *Proc. ACCV*, 2016, pp. 104–120.
[59] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.
[60] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4998–5006.

SAVINA JASSICA COLACO received the master's (Master in Technology) degree in computer science and engineering from CHRIST (Deemed to be University), Bengaluru, India, in 2019. She is currently pursuing the Ph.D. degree in electronic and electrical engineering with Kyungpook National University (KNU), Daegu, South Korea. In 2019, she joined the Intelligent Signal Processing Laboratory (ISPL), KNU, as a Ph.D. Degree Research Student. Her main research interests include computer vision, deep learning, and autonomous vehicles.

DONG SEOG HAN (Senior Member, IEEE) received the B.S. degree in electronic engineering from Kyungpook National University (KNU), Daegu, South Korea, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1989 and 1993, respectively. From 1987 to 1996, he was with Samsung Electronics Company Ltd., where he developed the transmission systems for QAM HDTV and Grand Alliance HDTV receivers. Since 1996, he has been a Professor with the School of Electronics Engineering, KNU. He was a Courtesy Associate Professor with the Department of Electrical and Computer Engineering, University of Florida, in 2004. He was the Director with the Center of Digital TV and Broadcasting, Institute for Information Technology Advancement, from 2006 to 2008. He is the Director at the center for ICT and Automotive Convergence, KNU. His main research interests include intelligent signal processing and autonomous vehicles.