3-D Facial Landmarks Detection for Intelligent Video Systems

Van-Thanh Hoang, *Student Member, IEEE*, De-Shuang Huang, *Senior Member, IEEE*, and Kang-Hyun Jo, *Senior Member, IEEE*

***Abstract*—Facial landmark detection is a fundamental research topic in computer vision that is widely adopted in many applications. Recently, thanks to the development of convolutional neural networks, this topic has been largely improved. This article proposes facial-landmark detector, which is based on a state-of-the-art architecture for landmark localization called stacked hourglass network, to obtain accurate facial landmark-points. More specifically, this article uses residual networks as the backbone instead of a 7×7 convolution layer. Additionally, it modifies the hourglass modules by using the residual-dense blocks in the mainstream for capturing more efficient features and the 1×1 convolution layers in the branch streams for reducing the model size and computational time, instead of the original residual blocks. The proposed architecture also enhances the features from modified hourglass modules with finer-resolution features via a lateral connection to generate more accurate results. The proposed network can outperform other state-of-the-art methods on the AFLW2000-3D dataset and the LS3D-W dataset, the largest three-dimensional (3-D face) alignment dataset to date.**

***Index Terms*—Convolution block, convolutional neural network (CNN), facial landmarks, stacked hourglass.**

I. Introduction

RECENTLY, many works have shown fantastic results on even the most challenging computer vision tasks, thanks to the advent of convolutional neural networks (CNNs) and the development of large datasets. This article pays attention to the facial landmark detection task, which is also known as face alignment.

Facial landmark detection or face alignment is one of the most heavily researched topics in computer vision over the last decades. It is the work of detecting the locations of facial landmark-points, like boundaries, eyes, elbows, and mouth, in a video or an image. As mentioned in [1], accurate facial landmark detection can improve the efficiency of many intelligent video systems, such as surveillance, people reidentification, facial animation, face modeling, and eye center localization [2].

Before the advent of CNNs, techniques based on hand-crafted features have been adopted for the task of landmark localization. For example, human pose estimation applications were mainly based on pictorial structures [3] and/or sophisticated extensions [4] due to their power of modeling large appearance changes and accommodating a wide spectrum of human poses.

Such methods have not been able to achieve the high performance exhibited by the cascaded regression methods [5]–[8]. These methods are recognized to crumble in cases of bad initialization and large (and unfamiliar) facial poses when there is a large in-plane rotation or a significant number of self-occluded landmarks.

Lately, fully convolutional neural network (FCNN) architectures [9] based on score-maps (also known as heat-maps) have revolutionized the human key points detection task [10]–[14] to produce accurate results, even for very challenging datasets [15]. Because the estimation of facial landmark and human key points problems are quite similar, such methods can be readily applied to the problem of facial landmark detection.

Nowadays, the task of predicting three-dimensional (3-D) facial landmark annotation is more interesting than two-dimensional (2-D) annotation. One main reason is that 2-D facial landmark annotation is not always semantically consistent and hardly preserves the 3-D structure of the human face, especially for profile views. Conversely, 3-D annotation preserves the correspondence across face poses. In this article, 3-D facial landmark annotation refers to 2-D projection of the actual 3-D landmark. But it is called the 3-D facial landmark for simplicity. The full 3-D facial landmark can be generated from this 3-D facial landmark by adding an extra network to predict the depth, as in [16].

This article detects the landmark-points of all faces in an input, which can be an image or a frame of video, by following the top-down approach. First, a face detector, such as Dlib [17], multitask cascaded CNN (MTCNN) [18], or Faster R-CNN (region-based CNN) [19], is used to detect the bounding boxes of faces inside the input. Then, the input is cropped based on these bounding boxes to have many cropped images, each image is corresponding to a bounding box (or a face). After that, for each cropped image, the score-maps of all landmark-points of the

Manuscript received March 27, 2019; revised September 23, 2019, October 28, 2019, and December 5, 2019; accepted December 26, 2019. Date of publication January 14, 2020; date of current version October 23, 2020. This work was supported by the grant of University of Ulsan. Paper no. TII-19-1085. *(Corresponding author: Kang-Hyun Jo.)*

V.-T. Hoang is with the Graduate School of Electrical Engineering, University of Ulsan, Ulsan 44610, South Korea (e-mail: thanhhv@islab.ulsan.ac.kr).

D.-S. Huang is with the Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: dshuang@tongji.edu.cn).

K.-H. Jo was with Tongji University, Shanghai, China, during his sabbatical leave in 2019 and now with University of Ulsan, Ulsan, South Korea (e-mail: acejo@ulsan.ac.kr).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2020.2966513

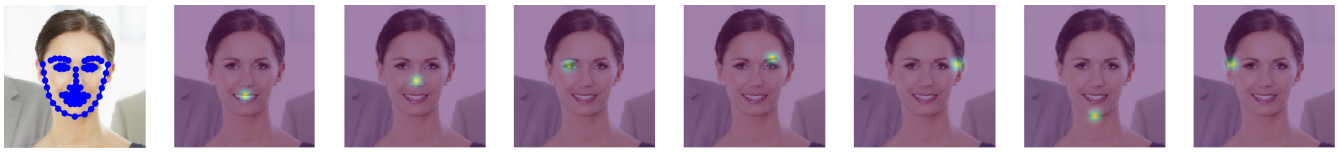

Fig. 1. Example output of the proposed facial-landmark detector. The first left image shows the facial landmark-points generated by the max activations across score-maps. The others show some sample score-maps of facial landmark-points (with the original image behind). From left to right: upper lip, nose, right eye, left eyebrow, left jaw, chin, and right jaw.

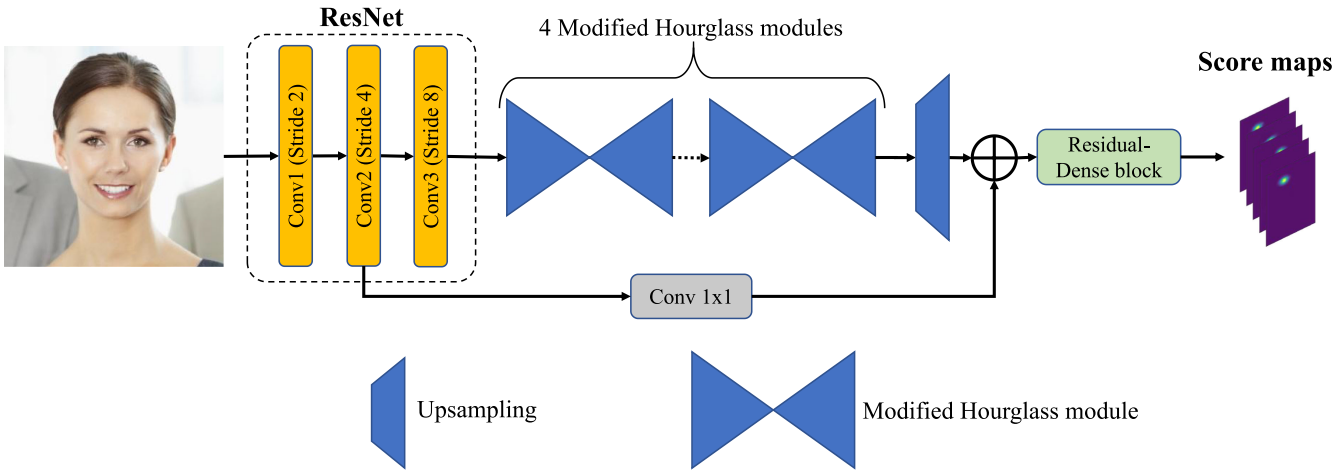

Fig. 2. Architecture of the proposed facial-landmark detector. It uses ResNet-50 as backbone, combines with four modified hourglass modules and one upsampling step to produce the score-maps for every facial landmark-points of the input face.

faces inside are generated by using the proposed facial-landmark detector. Some samples of generated score-maps are shown in Fig. 1. Finally, the landmarks of faces are aligned by the max activations across the score-maps and offset of bounding boxes on the original input image or frame.

The proposed facial-landmark detector is based on a state-of-the-art architecture for landmark localization called stacked hourglass network [11]. More specifically, as illustrated in Fig. 2, this article has contributed in the following three points.

1) It uses residual networks (ResNet) [20] as the backbone instead of a 7×7 convolution layer.
2) It modifies the hourglass modules by using the residual-dense blocks in the mainstream for capturing more efficient features and 1×1 convolution layers in the branch streams for reducing the model size and computational time, instead of the original residual blocks.
3) It also enhances the features from modified hourglass modules with finer resolution features via a lateral connection to generate high-accurate score-maps.

Based on these improvements, the proposed method achieves state-of-the-art performances on the AFLW2000-3D dataset and the large-scale 3-D faces in-the-wild (LS3D-W) dataset, the largest 3-D face alignment dataset to date.

II. Related Work

This section briefly introduces the related works on face detection, face alignment under two main categories (deep-learning-based methods and hand-crafted-features-based), convolution block design, and landmark localization.

A. Face Detection

Many face detection methods have been proposed [18], [19], [21]–[23]. Among them, MTCNN [18] and Faster R-CNN [19] are very well-known. Guided by the R-CNN family [24], Faster R-CNN [25] was originally proposed for object detection task. It consists of two steps in general. First, it generates the bounding-box proposals based on predefined anchors that only consider the objectness classification. After that, it crops the feature maps and then simultaneously detects the kind of inside object and refines the box proposals to obtain better final results. This research, for the face detection part, uses a Faster R-CNN detector with a ResNet-50 [20] backbone.

B. Hand-Crafted-Features-Based Methods for Face Alignment

The tree structure part model [26] used a deformable part-based model to model the face shape in a mixture of tree models for doing parallel detection, landmark localization, and pose estimation of faces in the image. Kamezi and Sullivan [27] introduced a real-time approach for aligning the facial landmarks with an ensemble of regression trees. The hand-crafted features like SIFT were also used by cascade regression-based methods [8], [28], [29] to capture the appearance of the face. However, those methods were unable to find out models for unrestricted faces in extreme poses. On the other hand, statistical methods, such as constrained local models (CLM) [30] and active appearance models [31] used HOG and SIFT [32] to perform key points estimation by maximizing the confidence of key-points positions in the input. Moreover, Asthana *et al.*

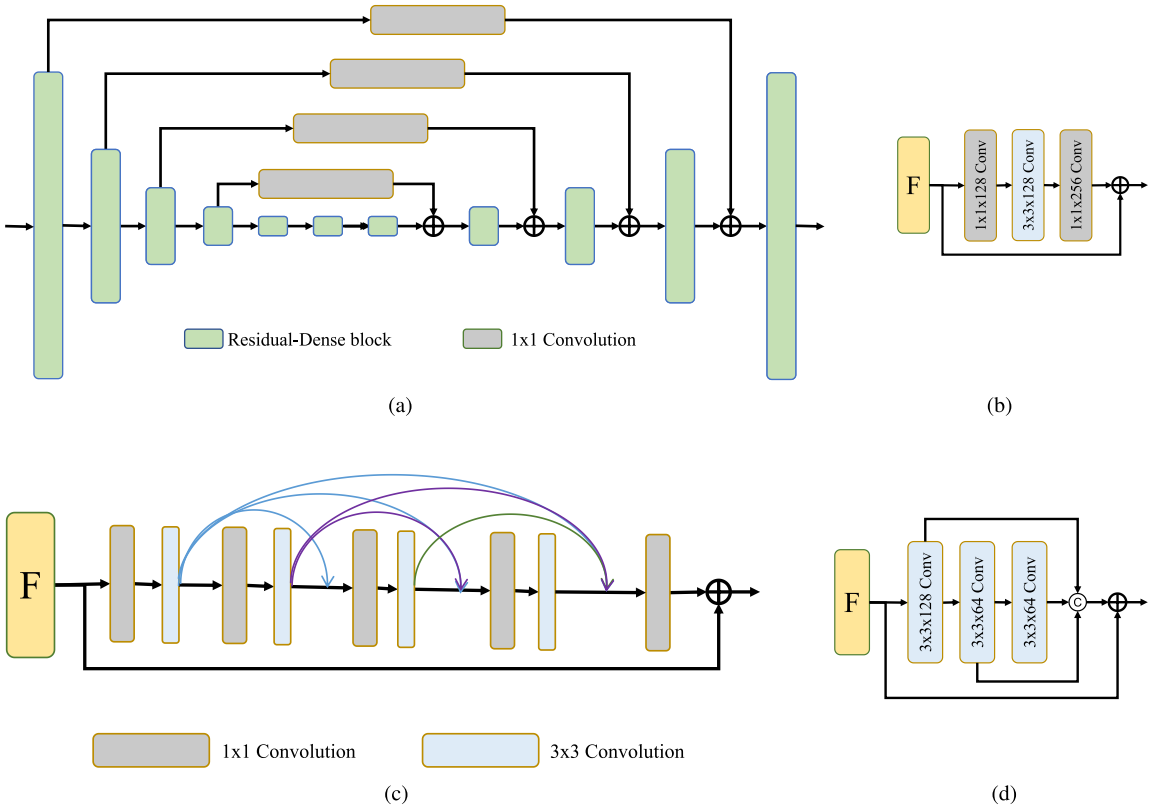

Fig. 3. Architecture of the modified hourglass module, original residual, residual-dense, and binarized blocks. (a) Modified hourglass module. (b) Original residual block. (c) Residual-dense block. (d) Binarized block.

[33] adopted linear regression in the CLM with a dictionary of the probability response maps.

C. Deep-Learning-Based Methods for Face Alignment

In [34], Sun *et al.* adopted a cascade of CNNs to regress the facial key-points locations. The work in [18] suggested a multitask learning system for attribute classification and facial landmark localization. The pose invariant face alignment method [35] used deep cascade regressors to generate the coefficients of a 3-D to 2-D projection matrix. This method was extended in [36] using CNNs to directly learn the dense 3-D coordinates. Merget *et al.* [37] detected the facial landmarks via a fully convolutional local-global context network. Zhu *et al.* [6] proposed the 3D dense face alignment (3DDFA) method to model the depth of the face, then fitted a dense 3-D model to the given input face image via a CNN. Bulat and Tzimiropoulos [38] modified the stacked hourglass networks [11] by using the Binarized block to reduce computational time but still achieved better results.

D. Convolution Block Design

He *et al.* [20] proposed the residual block with skip connections to create a residual mapping. The architecture of a residual block is shown in Fig. 3(b). As illustrated, after some plain convolution layers, the next layer gets the addition of the output of the preceding layer and the layer which comes before that as its input. Bulat and Tzimiropoulos [16] introduced an improved version of the residual block that has a hierarchical parallel and multiscale structure to improve the performance and efficiency called a Binarized block, as shown in Fig. 3(d). This block is designed to increase the receptive field size, improve gradient flow, and have (almost) the same number of parameters as the origin. Huang *et al.* [39] proposed another design called a dense block, which connects each layer to every other layer in a feedforward fashion. For each layer, the feature-maps of all preceding layers are used as inputs, and its feature-map is also used as input of all subsequent layers. This article introduces another block architecture called the residual-dense block, which is a residual block with a dense block inside for capturing more efficient features. The detailed architecture of this block is described in Section III-D.

E. Landmark Localization

Many CNN architectures, which based on the FCNN [9], are proposed for key points localization task, such as [11] and [38]. A state-of-the-art architecture is the stacked hourglass network [11]. It generates the score-maps for every key point by using a stack of eight hourglass modules. This article proposes the modified hourglass module, which uses the residual-dense blocks in the mainstream for capturing more efficient features and 1×1 convolution layers in the branch streams for reducing the model size and computational time, instead of the original residual blocks.

III. Proposed Approach

The proposed method adopts the top-down approach. From a given input image or frame, the Faster R-CNN face detector with ResNet-50 [20], as the backbone is used to generate the bounding boxes for all faces inside the input. Then, the input is cropped based on these bounding boxes to have many cropped images, each image is corresponding to a bounding box (or a face). After that, for each cropped image, the score-maps of all landmark-points of the faces inside is generated by using the proposed facial-landmark detector. Finally, the landmarks of faces are aligned by the max activations across these score-maps and the offset of bounding boxes on the original input image or frame.

A. Deep Residual Network

When the network goes deeper and deeper, the accuracy can be saturated and then degrades rapidly. Adding more layers to the model leads to lower accuracy. This problem is called the degradation problem. The ResNet is proposed by He *et al.* [20] to overcome this. Their solution is to add the skip connections to create a residual mapping. The architecture of the original residual block is shown in Fig. 3(b).

B. Facial-Landmark Detector

Fig. 2 shows the architecture of the proposed facial-landmark detector. Unlike the original stacked hourglass networks [11] or its variants [38], [40], which just use a 7×7 convolution layer at the beginning followed by multiple hourglass modules at stride $= 4$ (the resolution is 4×4 times lower than the input), this proposed network uses a backbone of three first blocks of ResNet-50 [20] to extract features of the input cropped image at stride $= 8$. Then, they are passed through four modified hourglass modules. Inspired by an excellent idea of U-net [41], the proposed network enhances the features from modified hourglass modules with finer resolution features via a lateral connection. More specifically, after hourglasses, it upsamples the spatial resolution of feature maps by a factor of 2 (using bilinear upsampling). The upsampled map is then merged with the corresponding bottom-up map (which undergoes a 1×1 convolution layer to fit channel dimension) by elementwise addition. Finally, it has an additional residual-dense block and a 1×1 convolution layer to output the score-maps for all key points.

The most expensive computational aspect of this system is the hourglass modules. In case of the original stacked hourglass networks [11] or its variants [38], [40], the hourglass modules have stride $= 4$ resolution, while in the proposed network, they have stride $= 8$. This means the proposed facial-landmark detector can be two or three times faster in comparison with those methods.

The upsampling step helps the proposed network generate the stride $= 4$ resolution score-maps with higher quality for landmark-points of the face inside. Some samples of generated score-maps are shown in Fig. 1.

C. Modified Hourglass Module

Fig. 3(a) illustrates the architecture of the modified hourglass module. It replaces the original residual blocks with residual-dense blocks in the mainstream to capture more efficient information. Since the residual-dense block is quite heavy in computation and size, it uses 1×1 convolution layers in the branch streams to reduce the model size and computational time.

D. Residual-Dense Block

Suppose a network has B blocks, each of them adopts a nonlinear transformation $H_b(\cdot)$, where b indexes the block. The output of the bth block is denoted as x_b.

In the traditional convolutional networks, the input of the $(b+1)$th block is the output of the bth block, which results in the following block transition:

$$x_b = H_b\left(x_{b-1}\right). \tag{1}$$

The residual block adds a residual connection that bypasses the $H_b(\cdot)$ with an identity function

$$x_b = H_b\left(x_{b-1}\right) + x_{b-1}. \tag{2}$$

The dense block uses the direct connections from any layer to all subsequent layers in a block. Consequently, the ℓth layer of the bth block receives the outputs of all preceding layers in the same block $x_b^0, \ldots, x_b^{\ell-1}$ as input calculated as

$$x_b^\ell = H_b^\ell\left(\left[x_b^0, x_b^1, \ldots, x_b^{\ell-1}\right]\right) \tag{3}$$

where $[x_b^0, x_b^1, \ldots, x_b^{\ell-1}]$ denotes the concatenation of the outputs of layers $0, 1, \ldots, \ell - 1$.

The residual-dense block is based on the concepts of the residual block and dense block. This block is similar to the dense block, except it also has a skip-connection that bypasses all L layers in the block with an identity function

$$x_b = H_b\left(\left[x_b^0, x_b^1, \ldots, x_b^{L-1}\right]\right) + x_{b-1}. \tag{4}$$

Fig. 3(c) illustrates the architecture of the residual-dense block used in the proposed network. The difference between this and the block described in [42] is before every 3×3 convolution layer, there is a 1×1 convolution layer to shrink the number of channels, which can reduce the model size and computational time.

IV. Experiments

A. Dataset

The proposed network is trained on the 300W-LP-3D datasets [6] and evaluated on two very challenging datasets for large 3-D face alignment. They are the LS3D-W dataset [16], the largest 3-D face alignment dataset to date, and the AFLW2000-3D dataset [6]. An important item to note is the 3-D annotations here are the 2-D projections of the actual 3-D facial landmarks. The datasets just call them 3-D facial landmarks for simplicity.

1) 300W-LP-3D: 300W-LP-3D is a synthetical dataset introduced in [6]. It is obtained by using the profiling method of [6] to render the faces of the 300-W dataset [43] into larger poses, ranging from $-90°$ to $+90°$. The dataset contains 61 225 images providing 3-D landmarks annotations.

TABLE I
MEAN AND STD OF THE NME (%) AND AUC (%) SCORES OF THE PROPOSED NETWORKS WITH DIFFERENT KINDS OF CONVOLUTION BLOCKS: RESIDUAL-DENSE, ORIGINAL RESIDUAL [20], AND BINARIZED [38]

Dataset	Residual-Dense		Residual		Binarized	
	AUC	Mean (StD)	AUC	Mean (StD)	AUC	Mean (StD)
Menpo-3D	**71.97**	**1.97 (1.72)**	69.83	2.18 (2.50)	70.49	2.09 (2.05)
300W-3D	**78.89**	**1.43 (0.42)**	77.07	1.56 (0.45)	77.53	1.53 (0.50)
AFLW2000-3D-Re	**74.22**	**1.86 (2.46)**	72.96	1.98 (2.84)	73.39	1.92 (2.54)

Bold texts denote outperformed scores.

2) LS3D-W: LS3D-W [16] is the largest 3-D face alignment dataset to date. This dataset is constructed by reannotating the 3-D landmarks for: the 300-W test set [43] (600 images), the Menpo dataset [44] (8955 images), and the 300-VW dataset [45] with a train set (95 217 images) and three test categories: Category A (62 643 images), Category B (32 872 images), and Category C (27 245 images). It should be noted that some images (especially from Category C) of 300-VW-3D have very low-resolution/poor quality faces. In total, it contains approximately 230 000 validation images.

3) AFLW2000-3D: AFLW2000-3D [6] is a dataset provided along with 300W-LP-3D. AFLW2000-3D is built by reannotating the first 2000 images of the annotated facial landmarks in the wild (AFLW) dataset [46] with 68 3-D landmarks. The faces in this dataset contain various expressions and illumination conditions with large-pose variations (yaw from $-90°$ to $+90°$). This dataset can be divided into three subsets: 1306 samples from $0°$ to $30°$, 462 samples from $30°$ to $60°$, and 232 samples in $[60°, 90°]$.

4) AFLW2000-3D-Reannotated: This dataset is provided by the authors in [16] along with the LS3D-W dataset. The AFLW2000-3D dataset is reannotated to make the ground truth more accurate because for faces with difficult poses, the method of [6] does not produce accurate landmarks.

B. Implementation Details

The proposed network is implemented via the MXNet framework [47]. It is trained using the Adam [48] optimizer implemented by MXNet on a server with an AMD Ryzen 7 3.60 GHz CPU, 32-GB RAM, and 2 NVIDIA 1080Ti GPU devices for 50 epochs. The learning rate is 4e$^-$5 and then decreased ten times at epochs of 20 and 35, respectively. The parameters are initialized by Xaviers initializer [49]. The other settings are weight decay of 0.0001, momentum of 0.9, and batch size of 32.

1) Data Augmentation: The proposed network is trained with random augmentation: color jittering, scale noise (from 0.5 to 1.2), blurring, flipping, and rotation (from $-50°$ to $+50°$).

2) Training Loss: The proposed network is trained to generate a set of score-maps, one for each landmark-point [50]. Where output centered at around the correct key-point location is placed in a Gaussian distribution and trained using the Sigmoid cross-entropy pixelwise loss, which is typically introduced for detection tasks [51]. This loss is defined as

$$\text{Loss}_{\text{Sigmoid}} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{W}\sum_{j=1}^{H} p_{kij}\log\hat{p}_{kij} + (1-p_{kij})\log(1-\hat{p}_{kij}) \quad (5)$$

TABLE II
MEAN AND STD OF NME (%) AND AUC (%) SCORES OF THE PROPOSED NETWORKS WITH TWO KINDS OF LOSSES: L2 LOSS AND SIGMOID LOSS

Dataset	L2 loss		Sigmoid loss	
	AUC	Mean (StD)	AUC	Mean (StD)
Menpo-3D	71.65	2.02 (2.08)	**71.97**	**1.97 (1.72)**
300W-3D	78.24	1.49 (0.69)	**78.89**	**1.43 (0.42)**
AFLW2000-3D-Re	74.20	1.90 (2.80)	**74.22**	**1.86 (2.46)**

Bold texts denote outperformed scores.

where K is the number of landmark-points ($K = 68$ in this article); W and H are the weight and height of the score-maps, respectively; $\hat{p}_{kij}$ denotes the predicted score-map output of the kth landmark-point at the output pixel location (i, j); and p_{kij} is the corresponding ground truth at the same location.

3) Evaluation Metrics: Similar to [6] and [38], this article uses normalized mean error (NME) and area-under-the-curve (AUC) scores as the metrics. In the case of the NME, this article analysis both Mean and standard deviation (StD) values (smaller is better). The NME is defined as

$$\text{NME} = \frac{1}{N}\sum_{n=1}^{N}\frac{\|y_n - \hat{y}_n\|_2}{d_n} \quad (6)$$

where N is the number of faces in the dataset; d_n is the square-root of the ground truth bounding box, which is computed as $d_n = \sqrt{w_{\text{bbox}} * h_{\text{bbox}}}$; y_n denotes the ground truth landmark for the given nth face; and $\hat{y_n}$ is the corresponding prediction landmark. The $\|\cdot\|_2$ denotes the ℓ_2-norm of the matrix. The AUC score (higher is better) is calculated based on the NME with threshold of 7%.

C. Ablation Study

1) Different Kinds of Convolution Blocks: Table I lists the scores of the proposed architecture with different kinds of convolution blocks (the original residual [20], binarized [38], and residual dense) on three datasets (Menpo-3D, 300-W-3D, and AFLW-2000-3D-Reannotated). It is easy to see that the network using residual-dense block achieves better scores on both AUC and NME (Mean and StD) than the others since the residual-dense block can capture more effective information.

2) Different Loss Function: Besides the Sigmoid cross-entropy pixelwise loss mentioned in Section IV-B, there is another kind of loss that is widely adopted in landmark localization task called L2 pixelwise loss [50], which is defined as

$$\text{Loss}_{L2} = \frac{1}{K}\sum_{k=1}^{K}\|Y_k - \hat{Y_k}\|_2 \quad (7)$$

TABLE III
MEAN AND STD OF NME (%) AND AUC (%) SCORES OF THE STATE-OF-THE-ART METHODS AND PROPOSED NETWORK ON THE LS3D-W DATASETS

Dataset	300VW-3D-CatA		300VW-3D-CatB		300VW-3D-CatC		Menpo-3D		300W-3D	
	AUC	Mean (StD)	AUC	Mean (StD)	AUC	Mean (StD)	AUC	Mean (StD)	AUC	Mean (StD)
3DDFA [6]	56.51	3.29 (2.83)	55.68	3.18 (**2.01**)	39.24	4.88 (3.99)	49.93	3.92 (3.30)	53.15	3.62 (3.13)
3D-FAN [38]	69.34	2.36 (3.75)	70.54	2.31 (3.93)	50.05	4.17 (6.03)	65.54	2.38 (**1.27**)	**81.09**	**1.27 (0.37)**
SAT2-CAB-3D [52]	63.56	2.73 (3.02)	66.36	2.61 (3.45)	41.02	4.62 (3.85)	63.47	2.70 (2.86)	71.02	2.06 (2.32)
HG2-CAB-3D [52]	64.18	2.67 (2.64)	66.86	2.49 (2.75)	40.35	4.54 (3.26)	64.19	2.63 (2.52)	70.67	2.02 (0.85)
Proposed	**73.52**	**1.93 (2.12)**	**74.63**	**1.94** (2.70)	**61.78**	**2.93 (3.48)**	**71.97**	**1.96** (1.72)	78.89	1.43 (0.42)

The results of 3DDFA [6], 3D-FAN [38], SAT2-CAB-3D [52], and HG2-CAB-3D [52] are obtained by running the public source code provided by the original authors. **Bold** texts denote outperformed scores.

TABLE IV
MEAN AND STD OF NME (%) AND AUC (%) SCORES OF THE STATE-OF-THE-ART METHODS AND PROPOSED NETWORK ON THE AFLW2000-3D AND AFLW2000-3D-REANNOTATED DATASETS

Method	AFLW2000-3D					AFLW2000-3D-Reannotated				
	AUC	$[0°, 30°]$	$[30°, 60°]$	$[60°, 90°]$	Mean (StD)	AUC	$[0°, 30°]$	$[30°, 60°]$	$[60°, 90°]$	Mean (StD)
CFSS [29]	-	4.77	6.71	11.79	7.76 (3.63)	-	-	-	-	-
MDM [28]	-	4.85	5.92	8.47	6.41 (1.86)	-	-	-	-	-
3DDFA [6]	-	4.11	4.38	5.16	4.55 (**0.54**)	-	-	-	-	-
3DDFA + SDM [6]	-	3.43	4.24	7.17	4.94 (-)	-	-	-	-	-
3D-FAN [38]	-	2.47	3.01	4.31	3.26 (-)	-	-	-	-	-
CMHM [52]	-	2.36	**2.80**	**4.08**	3.08 (-)	-	-	-	-	-
DHM [53]	-	2.75	4.21	6.91	4.62 (-)	-	2.28	3.10	6.95	4.11 (-)
DHM + RHG [53]	-	2.52	3.21	5.76	3.85 (-)	-	2.25	3.05	4.21	3.17 (-)
3DDFA* [6]	49.90	3.11	4.18	5.52	3.68 (2.71)	57.32	2.58	3.48	5.03	3.13 (2.48)
3D-FAN* [38]	57.66	2.51	3.27	4.46	2.95 (1.47)	72.69	1.85	1.84	**2.24**	1.91 (**1.77**)
SAT2-CAB-3D* [52]	56.72	2.53	3.55	4.79	3.07 (1.96)	68.39	1.79	2.55	3.83	2.25 (1.91)
HG2-CAB-3D* [52]	57.71	2.50	3.33	4.78	3.01 (2.18)	69.58	1.75	2.31	3.76	2.17 (2.16)
Proposed	**62.47**	**2.29**	2.90	4.32	**2.71** (2.59)	**74.22**	**1.59**	**1.75**	3.16	**1.86** (2.46)

The results of 3DDFA [6], 3D-FAN [38], SAT2-CAB-3D [52], and HG2-CAB-3D [52] are also obtained by running the public source code provided by the original authors (denoted by *). The results of CFSS [29] and MDM [28] are obtained from [6]. **Bold** texts denote outperformed scores.

where $\hat{Y_k}$ denotes the predicted score-map of the kth landmark-point, and Y_k is the corresponding ground truth score-map. Here, the $\|\cdot\|_2$ denotes the ℓ_2-norm of the matrix. The way to generate ground truth score-maps is the same as Sigmoid loss.

Table II describes the scores of the proposed network trained with two losses on three datasets: Menpo-3D, 300-W-3D, and AFLW-2000-3D-Reannotated. The results indicate that these two losses offer similar performance for the proposed network in terms of AUC and Mean of NME. However, the network trained with L2 loss is much worse than the Sigmoid loss in terms of StD of NME.

The outputted score-maps can be considered as maps of the probability that a pixel corresponds to the position of a key point, with a small number of positions that have nonzero probability. Thus, similar to the classification task, the Sigmoid loss is better because the L2 loss gives too much emphasis to the zero-probability positions. That is why, similar to [38], the Sigmoid loss function is used in training the proposed network.

D. Quantitative Results

Table III describes the Mean and StD of NME and AUC scores of the state-of-the-art methods (3DDFA [6], 3D face alignment network (3D-FAN) [38], SAT2-CAB-3D [52], and HG2-CAB-3D [52]) and the proposed network on the datasets 300-VW-3D (three categories), 300-W-3D, and Menpo-3D, which are parts of the LS3D-W dataset. Because the authors of those works did not report their results for these datasets, this article validates those methods based on their public code.

As can be seen, the proposed network outperforms the state-of-the-art methods on the datasets except for the 300-W-3D dataset. However, the 300-W-3D dataset is quite small with just 600 images, while the others have thousands of images.

Table IV describes the Mean and StD of NME and AUC scores of the state-of-the-art methods and proposed network on the AFLW2000-3D and AFLW2000-3D-Reannotated datasets. Considering methods from [52], 3DDFA [6], and 3D-FAN [38], the authors reported their results on the AFLW2000-3D dataset; however, they have added some improvements in their public code, so this article reports the results obtained from both the article and running the public code.

As can be seen, the proposed network also outperforms the other methods, except for 3D-FAN. It has better AUC and Mean of NME scores, but worse in StD of NME when compared to 3D-FAN. A clear trend here is as the yaw angle increases, most of the methods begin to degrade. This may be due to a small number of training faces, which have a large yaw angle.

E. Qualitative Results

Some visual examples of 3-D facial landmarks detected by the proposed method are shown in Fig. 4. As illustrated, the detected and ground truth 3-D facial landmarks are similar in most cases. Unfortunately, the detected landmarks are mixed up in some complicated cases such as very low resolution, bad lighting, more than 90° yaw angle rotation, and/or complex face color.

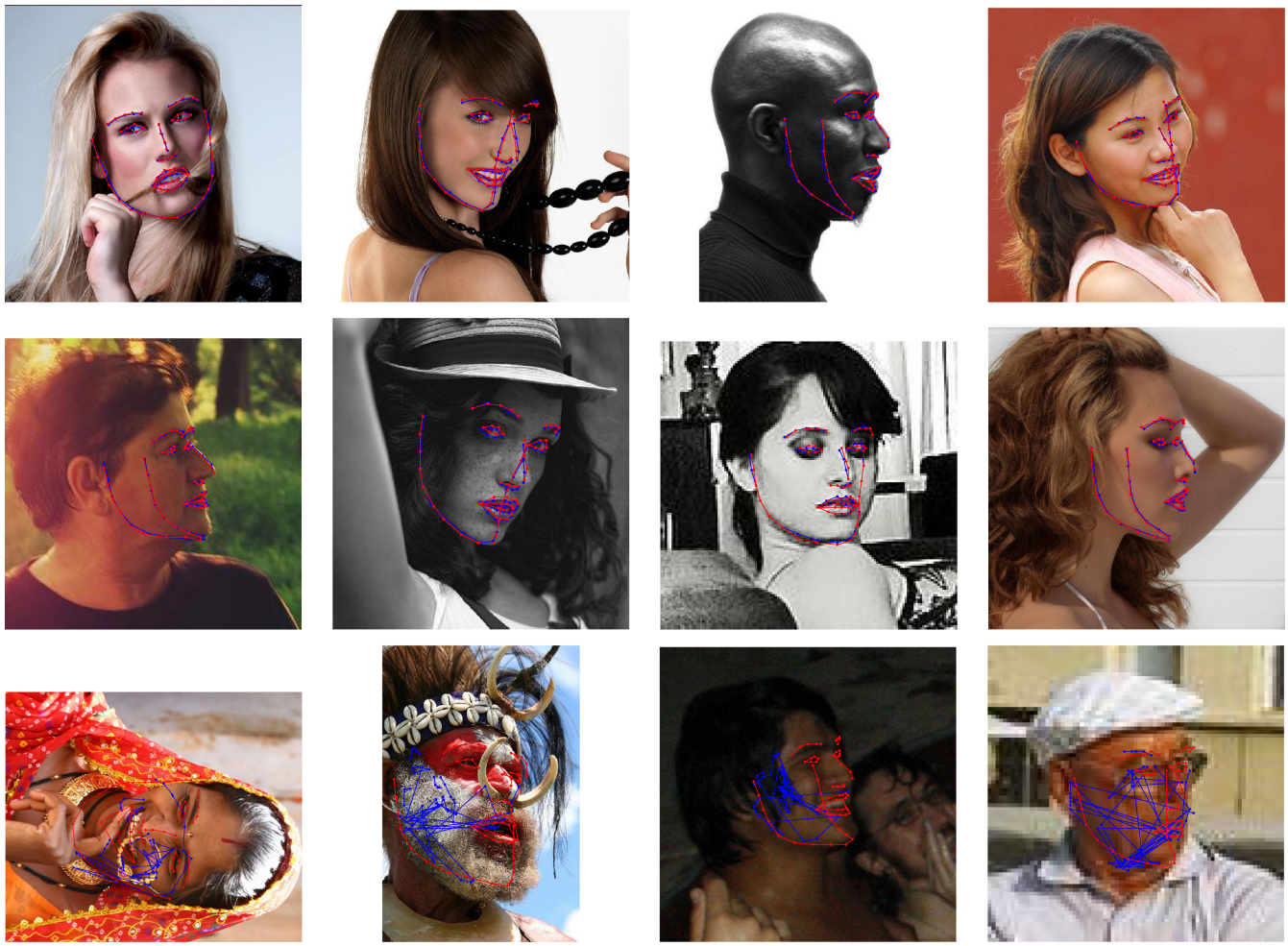

Fig. 4. Qualitative results of the proposed facial-landmark detector. Blue: detected landmarks. Red: ground truth landmarks. The top two rows are the examples of good results when the ground truth and detected facial landmarks are similar. The last row is the examples of the lowest accuracy results when detected landmarks are mixed up.

This may be due to the training dataset not containing many images with these complicated situations.

F. Runtime Analysis

Finally, this article evaluates the actual inference speeds of the methods on a computer mentioned in Section IV-B. The face detection part using Faster R-CNN (with 300 proposals and image size of 600×1000) takes 60 ms. For the cropped image, the proposed facial-landmark detector takes 60 and 40 ms; 3DDFA takes 5 and 3 ms; methods of [52] take 100 and 50 ms; and 3D-FAN takes 158 and 125 ms for images sizes of 256×256 pixels and 128×128 pixels, respectively.

As can be seen, the 3DDFA is the fastest, but it has much lower scores than the others. The proposed method is approximately $2.5\times$ faster than 3D-FAN, thanks to the modification mentioned above.

To obtain a faster system, there are some quicker face detectors, which can be used instead of Faster R-CNN, such as the Dlib [17] or MTCNN [18] face detector. Another way is replacing the ResNet-50 backbone with some smaller network architecture like MobileNet [54].

V. Conclusion

This article introduced the facial-landmark detector to detect the 3-D facial landmarks for faces in a video or an image. This network used the ResNet-50 as the backbone, followed by four modified hourglass modules. It modified the hourglass modules by using the residual-dense blocks in the mainstream for capturing more efficient features, as well as 1×1 convolution layers in the branch streams for reducing the model size and computational time, instead of the original residual blocks. It also enhanced the features from modified hourglass modules with finer resolution features via a lateral connection to generate higher-accuracy score-maps. The proposed method can achieve not only higher scores, but also faster speed compared with other state-of-the-art methods.

In the future, the Faster R-CNN and facial-landmark detector can be combined into one system since both of them have similar backbone, ResNet-50. Additionally, ResNet-50 can be replaced

with some lightweight architecture like MobileNet [54] to make the system smaller and faster so that it can be used in real-time applications.

REFERENCES

[1] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa, "The do's and don'ts for CNN-based face verification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 2545–2554.

[2] Z. Wang, H. Cai, and H. Liu, "Robust eye center localization based on an improved SVR method," in *Proc. Int. Conf. Neural Inf. Process.*, 2018, pp. 623–634.

[3] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[4] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 588–595.

[5] V.-T. Hoang, V.-D. Hoang, and K.-H. Jo, "An improved method for 3d shape estimation using cascade of neural networks," in *Proc. IEEE Int. Conf. Ind. Informat.*, 2017, pp. 285–289.

[6] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, Jan. 2019.

[7] V.-T. Hoang and K.-H. Jo, "3D human pose estimation using cascade of multiple neural networks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2064–2072, Apr. 2019.

[8] 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.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[10] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 34–50.

[11] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[12] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.

[13] V.-T. Hoang and K.-H. Jo, "Multi-person pose estimation with human detection: A parallel approach," in *Proc. Annu. Conf. IEEE Ind. Electron. Soc.*, 2018, pp. 3269–3272.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[15] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3686–3693.

[16] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1021–1030.

[17] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, no. 7, pp. 1755–1758, 2009.

[18] 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.

[19] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 650–657.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[21] Q. Huang, C. K. Jia, X. Zhang, and Y. Ye, "Learning discriminative subspace models for weakly supervised face detection," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 2956–2964, Dec. 2017.

[22] S. Jin, D. Kim, T. T. Nguyen, D. Kim, M. Kim, and J. W. Jeon, "Design and implementation of a pipelined datapath for high-speed face detection using FPGA," *IEEE Trans. Ind. Informat.*, vol. 8, no. 1, pp. 158–167, Feb. 2012.

[23] P. Vadakkepat, P. Lim, L. C. De Silva, L. Jing, and L. L. Ling, "Multimodal approach to human-face detection and tracking," *IEEE Trans. Ind. Electron.*, vol. 55, no. 3, pp. 1385–1393, Mar. 2008.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[26] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879–2886.

[27] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[28] 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.*, 2016, pp. 4177–4187.

[29] 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.*, 2015, pp. 4998–5006.

[30] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognit.*, vol. 41, no. 10, pp. 3054–3067, 2008.

[31] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[32] P. C. Ng and S. Henikoff, "Sift: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, 2003.

[33] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3444–3451.

[34] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3476–3483.

[35] A. Jourabloo and X. Liu, "Pose-invariant 3D face alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3694–3702.

[36] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3219–3228.

[37] D. Merget, M. Rock, and G. Rigoll, "Robust facial landmark detection via a fully-convolutional local-global context network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 781–790.

[38] A. Bulat and G. Tzimiropoulos, "Hierarchical binary CNNs for landmark localization with limited resources," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 343–356, Feb. 2020.

[39] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[40] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected u-nets for efficient landmark localization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 339–354.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[42] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.

[43] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, 2016.

[44] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 170–179.

[45] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic1, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 50–58.

[46] M. Kstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2144–2151.

[47] T. Chen *et al.*, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," in *Proc. Neural Inf. Process. Syst., Workshop Mach. Learn. Syst.*, 2015. [Online]. Available: https://github.com/apache/incubator-mxnet#reference-paper

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:main.html

[49] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[50] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.
[51] N. Zhang, E. Shelhamer, Y. Gao, and T. Darrell, "Fine-grained pose prediction, normalization, and recognition," in *Proc. Int. Conf. Learn. Representations Workshops*, 2016. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2016:main.html
[52] J. Deng, Y. Zhou, S. Cheng, and S. Zaferiou, "Cascade multi-view hourglass model for robust 3D face alignment," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 399–403.
[53] B. Sun, M. Shao, S. Xia, and Y. Fu, "Deep evolutionary 3D diffusion heat maps for large-pose face alignment," in *Proc. Brit. Mach. Vis. Conf.*, 2018. [Online]. Available: http://bmvc2018.org/programmedetail.html
[54] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

Van-Thanh Hoang (Student Member, IEEE) received the bachelor's degree in information technology and the master's degree in computer science from the Posts and Telecommunications Institute of Technology, Hanoi, Vietnam, in 2011 and 2013, respectively. He is currently working toward the Ph.D. degree in electrical and computer engineering with the Graduate School of Electrical Engineering, University of Ulsan, Ulsan, South Korea.

He was a Software Engineer with VCCorp, Vietnam until the middle of 2012 and then for Vivas company until the end of 2013. Since 2014, he has been a Lecturer with the Department of Engineering and Technology, Quang Binh University, Quang Binh, Vietnam. His research interests include pattern recognition, computer vision, bioinformatics, and machine learning.

De-Shuang Huang (Senior Member, IEEE) received the B.Sc. degree from Institute of Electronic Engineering, Hefei, China, in 1986, the M.Sc. degree from the National Defense University of Science and Technology, in 1989, Changsha, China, and the Ph.D. degree from Xidian University, Xian, China, in 1993, all in electronic engineering. During 1993 and 1997, he was a Postdoctoral Student, respectively, with the Beijing Institute of Technology, Beijing, China, and the National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China.

In September 2000, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences as the Recipient of Hundred Talents Program of CAS. In September 2011, he entered into Tongji University as a Chaired Professor. From September 2000 to March 2001, he was a Research Associate with Hong Kong Polytechnic University. From August to September 2003, he visited the George Washington University as a Visiting Professor, Washington DC, USA. From July to December 2004, he was the University Fellow with the Hong Kong Baptist University. From March 2005 to March 2006, he was a Research Fellow with the Chinese University of Hong Kong. From March to July 2006, he was a Visiting Professor with the Queens University of Belfast, U.K. In 2007, 2008, 2009, he was a Visiting Professor with Inha University, South Korea. He is currently the Director of the Institute of Machines Learning and Systems Biology, Tongji University, Shanghai, China. He has authored or coauthored more than 180 journal papers. His current research interest includes bioinformatics, pattern recognition, and machine learning.

Dr. Huang is currently the IAPR Fellow.

Kang-Hyun Jo (Senior Member, IEEE) received the Ph.D. degree in computer controlled machinery from Osaka University, Osaka, Japan, in 1997.

After a year of experience with ETRI as a Postdoctoral Research Fellow, he joined the School of Electrical Engineering, University of Ulsan, Ulsan, South Korea. He is currently serving as the Faculty Dean with the School of Electrical Engineering, University of Ulsan, Ulsan, South Korea. His research interests include computer vision, robotics, autonomous vehicle, and ambient intelligence.

Dr. Jo has served as the Director or an AdCom Member with the Institute of Control, Robotics and Systems, The Society of Instrument and Control Engineers, and the IEEE IES Technical Committee on Human Factors Chair, AdCom Member, and the Secretary until 2019. He has also been involved in organizing many international conferences, such as International Workshop on Frontiers of Computer Vision, International Conference on Intelligent Computation, International Conference on Industrial Technology, International Conference on Human System Interactions, and the Annual Conference of the IEEE Industrial Electronics Society. He is currently an Editorial Board Member for international journals, such as the *International Journal of Control, Automation, and Systems* and *Transactions on Computational Collective Intelligence*.