IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 February 2023, accepted 18 March 2023, date of publication 27 March 2023, date of current version 30 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262247

RESEARCH ARTICLE

Facial Landmark, Head Pose, and Occlusion Analysis Using Multitask Stacked Hourglass

YOUNGSAM KIM, JONG-HYUK ROH, AND SOOHYUNG KIM
Information Security Research Division, Electronics and Telecommunications Research Institute, Daejeon 34129, South Korea

Corresponding author: Youngsam Kim (kimzt@etri.re.kr)

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00321, Privacy preserving self-controlled decentralized identity management and security technology in 5G service environment).

ABSTRACT In this study, we proposed a multitask network architecture for three attributes, landmark, head pose, and occlusion, from a face image. A 2-stacked hourglass with three task-specific heads is the proposed network architecture. We also designed three auxiliary components for the network. First is the feature pyramid fusion module, which plays a crucial role in facilitating contextual information from various receptive fields. Second is the interlevel occlusion-aware fusion module, which explicitly fuses intermediate occlusion prediction between subnetworks. The third is the gimbal-lock-free head pose head, which outputs a rotation matrix from a 6D rotation representation. We conducted an ablative study of these auxiliary components to determine their impacts on the network. Additionally, we introduced the landmark heatmap scaling approach to avoid falling local minima. We trained the proposed network with a 300W-LP dataset for landmark and head pose and a C-CM dataset for occlusion. Then, we fine-tuned the network using the 300W or WFLW dataset, instead of the 300W-LP dataset for the landmark task. This 2-stage training method contributes to enhancing the landmark detection accuracy and that of other tasks. In the experiments, we assessed the performance of the proposed network on eight test datasets using task-specific metrics. The results show that the proposed network achieved competitive performance across all the datasets and notably outperformed the state-of-the-art methods on AFLW2000 and Masked 300W datasets.

INDEX TERMS Landmark detection, head pose estimation, occlusion segmentation, multitask learning, deep neural networks, face analysis.

I. INTRODUCTION

For a face recognition system, facial landmark detection, head pose estimation, and occlusion segmentation tasks are challenging and actively researched problems. Face recognition systems outperformed humans with the advent of the deep convolutional neural network (CNN); however, in practice, it is still necessary to enhance the robustness to noises, such as pose, occlusion, or expression. If these types of noises are examined, noise-robust approaches such as face frontalization [1], [2], [3] or occlusion-aware approaches [4], [5], could be applied to the face recognition system. To analyze facial landmarks, head pose, and occlusion (FaceLPO), we can follow one of two methods: single-task learning (STL) and multitask learning (MTL). The STL method requires training three independent networks, one for each task. The network can be easily optimized and fine-tuned by focusing on a single task; however, it becomes less memory and computationally efficient if the number of tasks is increased. In contrast, the MTL method shares one network for multiple tasks, making it more memory and computationally efficient. However, optimizing the shared network can be more challenging.

In this study, we first introduce the MTL method to investigate FaceLPO. The network architecture and the locations of each task are two of the most crucial components of the MTL method. According to [6], the head pose is a global attribute and requires a low-resolution feature map, whereas landmark and occlusion tasks are position-sensitive and require a high-resolution feature map. We designed

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

multitask encoder-decoder architecture based on a stacked hourglass (HG) network [7] to satisfy these characteristics, with the head pose task located at the end of the encoder and landmark and occlusion tasks located at the end of the decoder.

Additionally, we proposed three auxiliary components to enhance the performance of each task. First, we proposed a feature pyramid fusion (FPF) module. Previous studies [8], [9], [10] demonstrated that contextual information helps enhance the accuracy of semantic segmentation tasks. Inspired by the previous studies, we reinforced the contextual information by fusing numerous feature maps with different resolutions in a decoder. Second, we added an interlevel occlusion-aware fusion (IOAF) module. A stacked HG network passes the outputs of the previous subnetwork to the next subnetwork. The IOAF module fuses the intermediate outputs, landmark heatmap, and occlusion mask to enhance occlusion awareness of the network. Third, we applied gimbal-lock-free 6D rotation representation instead of 3D Euler angles. Euler angles are not an optimal representation of the head pose due to the gimbal-lock problem [11]. We modify the head pose head to output 6D rotation parameters and convert it to a rotation matrix using Gram–Schmidt orthonormal process [12].

A publicly available training dataset that simultaneously supports FaceLPO labels is needed to jointly train the proposed network. However, we could not find that type of dataset. Instead, we selected 300W-LP [13] and C-CM [14] datasets. 300W-LP provides landmark and head pose labels and C-CM provides occlusion segmentation labels. We also adopt a 2-stage training approach. This contributes to enhancing the accuracy of all tasks by employing pretrained weights. Additionally, it mitigates the bias of landmark points, as mentioned in [15].

We assess our model for eight test datasets; 300W [16], WFLW [17], Masked 300W [18] for landmark detection, AFLW2000 [13], BIWI [19], AFLW2000-SO for the head pose estimation, COFW [20], and RealOcc-Wild [14] for occlusion segmentation. Masked 300W and AFLW2000-SO datasets are used to assess the occlusion robustness of our model in landmark detection and head pose estimation. We generated AFLW2000-SO by artificially applying simulated occlusions to the AFLW2000 dataset, inspired by Borghi et al. [21]. Our model achieved the best accuracy for AFLW2000, Masked 300W, and second-best for WFLW, COFW, and RealOcc-Wild. Furthermore, we conducted an ablative study for the proposed three auxiliary components. The study demonstrates that FPF and IOAF modules can enhance the accuracy for all tasks, but 6D rotation representation did not show the difference.

In summary, we first propose, LPONet, a multitask encoder-decoder network for FaceLPO tasks. This demonstrated competitive performance over the other SOTAs. We visualize our predictions for some images with occlusion or large pose in WFLW, AFLW2000, and RealOcc-Wild datasets, as shown in Figure 1. This study is organized as follows: Section II includes previous studies related to each facial landmark detection, head pose estimation, and occlusion segmentation. Section III describes the proposed architecture, three auxiliary components, and the loss function. In Section IV, we evaluate the proposed model and compare its performance with SOTAs. Finally, Section V concludes the study.

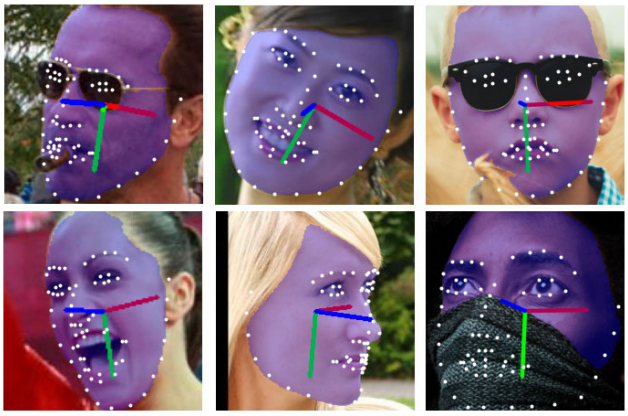


FIGURE 1. Visualization of simultaneous FaceLPO predictions on three different datasets. WFLW test set, AFLW2000, and RealOcc-Wild from left column.

II. RELATED WORKS

In this section, we review previous studies related to the MTL-based analysis of facial attributes, STL-based landmark detection, head pose estimation, and occlusion segmentation.

A. MULTITASK LEARNING

Numerous multitask methods [6], [22], [23], [24] have been used in face detection, alignment, and analysis problems. Zhang et al. [22] developed a 3-stage cascaded structured CNN and trained it with regression loss for the facial bounding box and five landmarks. To examine face alignment, the head pose, gender, age, and expression from a cropped face, Ranjan et al. [23] proposed a single multitask CNN. The network consists of two subnetworks; one is for subject-independent tasks and the other is for subject-dependent tasks. They trained all tasks simultaneously in an end-to-end manner. Furthermore, Ranjan et al. [24] proposed a HyperFace network for four tasks; face detection, alignment, pose estimation, and gender recognition. Valle et al. [6] proposed U-Net [25] based multitask architecture for a head pose, 68 landmarks, and visibility. These studies demonstrated that multiple face attributes can be effectively examined using a single model with the MTL method, and even achieved superior performances by developing synergy among correlated tasks. However, the MTL method for the FaceLPO tasks had not yet been investigated.

B. LANDMARK DETECTION

The landmark detection task is to determine the locations of the eyes, nose, mouth, and jaw from a face image. In detail, this task comprises two categories; 2D and 3D

landmarks. In this study, we only focus on 2D landmark detection. Recently, several studies [15], [17], [26], [27], [28], [29], [30], and [31] on 2D landmark detection have been performed. Bulat and Tzimiropoulos [15] proposed a heatmap regression method based on a stacked HG network. They pointed out the difference in ground-truth landmark points between datasets and proposed fine-tuning approach to reduce this labeling bias. Guo et al. [26] designed a landmark coordinate regressor employing MobileNet [32] block. In the benchmark test employing an ARM processor, they achieved 37 FPS for the default model and 140 FPS for a thin model.

Loss functions are also crucial parts when training neural networks. In either coordinates or heatmap regression, it is an issue that the error becomes too small when the network is trained. Thus, the network became early saturated. Wing loss [27] and adaptive wing loss [28] have been proposed to address this issue. Wing loss is designed to have a significant influence when the error gets smaller. The goal of adaptive wing loss is the same as wing loss, except that it is designed for heatmap regression. Adaptive wing loss applied a modified logarithm function to the foreground pixel and an L2 loss function to the background pixel. This causes the network to focus on the error in foreground pixels than the background, as training goes on.

Wu et al. [17] designed a boundary-aware face alignment framework that combines coordinate regression and heatmap regression methods. The framework first outputs boundary heatmap, instead of landmark heatmap and sends it to the encoder network, which outputs landmark coordinates. Hsu et al. [29] investigated the characteristics of heatmap and coordinates-based methods. Thus, they designed a hybrid loss function with pixel-wise classification loss and coordinate regression loss. Jin et al. [30] proposed a Pixel-in-Pixel network (PIPNet) based novel hybrid method. PIPNet is encoder-only architecture with a PIP regression head. The PIP regression head is attached to the medium-resolution layer that outputs both the coarse grid heatmap and offsets in each grid. Furthermore, they proposed a neighbor regression module that uses the interrelationship between a landmark point and some neighbors. Bulat et al. [31] identified quantization errors when encoding ground-truth heatmaps from coordinates. They addressed the problem using a continuous heatmap encoding approach and achieved the best landmark detection accuracy.

C. HEAD POSE ESTIMATION

Studies on head pose estimation include both RGB image- and depth-based approaches. Depth-based methods can be more accurate but require special cameras, which are not always available [33]. Therefore, we only focus on RGB image-based methods in this study.

Traditionally, head pose estimation has been handled by POSIT [34], which exploits correspondence between 3D head models and 2D landmarks. Recently, the deep learning method is widely employed to estimate the head pose. The training dataset is crucial for the deep learning method. Liu et al. [35] developed a 3D synthetic head pose dataset employing a rendering tool and proposed a CNN, which directly regresses Euler angles; pitch, yaw, and roll. Ruiz et al. [36] proposed a multi-loss method. The multi-loss combines a binned pose classification loss and a regression loss. Zhou and Gregson [37] extended the HopeNet [36] to deal with wide yaw. Zhang et al. [38] developed a feature decoupling module to decrease dependency among feature spaces for each angle. They also designed cross-category center loss to increase intra-class compactness and inter-category separability, simultaneously.

Based on a 2-stream network, Yang et al. [33] proposed a fine-grained structure aggregation network (FSA-Net). It aggregates two intermediate feature maps from each stream employing an attention mechanism. Each stream consists of 10 layers and each layer is a depthwise separable convolution layer. FSA-Net demonstrated competitive performance with previous SOTAs despite 1MB tiny model. Zhou et al. [12] mentioned that the discontinuity of 3D or 4D rotation representation for full range hinders optimizing neural networks. To address the issue, they proposed a new 6D continuous rotation representation. Hempel et al. [11] applied the 6D rotation representation to the head pose estimation task and showed superior accuracy.

D. OCCLUSION SEGMENTATION

A representative semantic segmentation network is a fully convolutional network (FCN) [39], which turns the classification network into the segmentation network. FCN is based on encoder-only architecture such as AlexNet [40] or VGGNet [41]. FCN selects and upsamples multiple low-resolution feature maps to get a high-resolution feature map. Then, pixel-wise classification is executed. Afterward, numerous researchers tried to obtain more accurate segmentation mask predictions. Chen et al. [42] and [43] pointed out that the deeper layer has abundant contextual information, but less spatial information, leading to the inaccurate boundary of the mask. They proposed to employ atrous convolution as decreasing pooling layers.

Some studies [8], [9], [10] also attempted to employ long-range context information. Zhao et al. [8] proposed a pyramid scene parsing network (PSPNet) with a pyramid pooling module, which uses context information from different sub-regions through multiple pooling layers with varying kernel sizes. Zhao et al. [9] proposed a point-wise spatial attention network (PSANet) and achieved better performance than PSPNet. Zhu et al. [10] designed an asymmetric non-local network with a non-local block [44] fused with a pyramid pooling module. With ResNet [45] backbone, it achieved better performance than PSANet.

Meanwhile, encoder-decoder architectures [25], [46], [47] have been investigated. U-Net [25] is designed to have a decoder which is symmetric to an encoder and won the ISBI

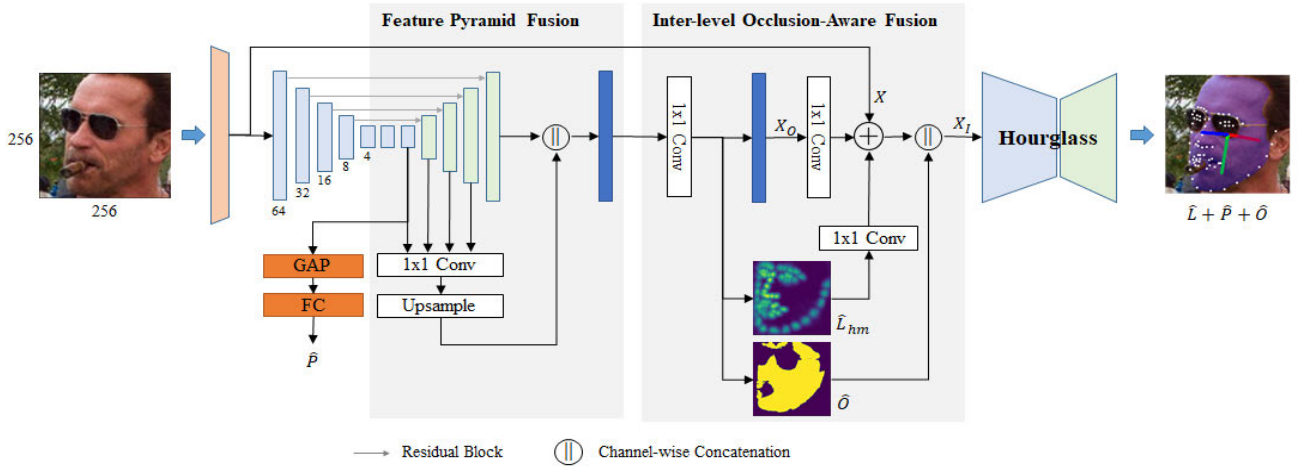


FIGURE 2. Overall network architecture of LPONet.

cell tracking challenge in 2015. DeconvNet [46] modified upsampling approach from bilinear interpolation to deconvolution. SegNet [47] proposed a memory-efficient max-unpooling approach for upsampling. Furthermore, there are some unique architectures; HRNet [48], which employs low- and high-resolution feature maps in parallel, and SegFormer [49] based on vision Transformer [50].

Face occlusion segmentation tasks [14], [51], [52], [53] employ the above semantic segmentation approaches. Particularly, Voo et al. [14] opened their face occlusion dataset, based on CelebAMask-HQ [54]. They obtained binary segmentation masks from original labels and manually corrected the wrong ones. They also classified the images as clean and occluded faces. Consequently, they synthesized face images with hands, COCO [55] objects, or random shapes. Yin and Chen [53] also synthesized a face occlusion dataset based on CelebAMask-HQ, but it is currently unavailable.

III. NETWORK ARCHITECTURE

In this section, we introduce our proposed multitask stacked HG network called LPONet. Furthermore, we describe three auxiliary components. First is the FPF module to enhance the performance of the landmark and occlusion tasks. The second is an IOAF module to enhance the occlusion robustness of all tasks. The third is the gimbal-lock-free 6D rotation representation head.

A. LPONET

The proposed network should have a structure appropriate for all three tasks, FaceLPO. According to previous studies, landmark detection and occlusion segmentation tasks are suitable for encoder-decoder architecture, which outputs high-resolution feature maps. However, the head pose task is suitable for encoder-only architecture, which outputs an embedding vector containing global context information. Thus, we selected a stacked HG network, which is a representative encoder-decoder architecture as a backbone, and located task-specific heads at the end of the encoder and decoder.

A stacked HG network consists of a stem block and n HG subnetworks. The stem block starts with a 7×7 convolutional layer with stride 2, followed by a residual block, max-pooling and two more residual blocks. The 256×256 input image is downsampled to a 64×64 feature map through the stem block. Then, HG subnetworks are repeated n times. The subnetwork has encoder and decoder parts. The encoder part encodes the 64×64 input feature map to the output 4×4 feature map through seven layers. The decoder part upsamples the encoded feature map to the 64×64 feature map. The encoder's feature maps are added to the decoder's corresponding feature maps by lateral skip connection when upsampling. It could compensate for the positional information lost by pooling. Each layer is composed of a residual block, the number of output channels C is 256, and the number of stacks n is 2.

As shown in Figure 2, the proposed LPONet is based on a 2-stacked HG with three task-specific heads. The head pose head applies global average pooling to the encoder's output followed by a fully connected (FC) layer, which outputs a 256D embedding vector. Then, one more FC layer is applied and outputs 3D head pose parameters $\hat{P}$. In the landmark head, it applies a 1×1 convolutional layer to the decoder's output and obtains $64 \times 64 \times 68$ landmark heatmap $\hat{L}_{hm}$. In the occlusion head, it applies a 1×1 convolutional layer to the decoder's output shared with the landmark head and obtains a $64 \times 64 \times 2$ occlusion mask $\hat{O}$. Landmark coordinates $\hat{L}$ can be obtained by applying channel-wise argmax to the $\hat{L}_{hm}$.

B. FPF

Landmark and occlusion analysis must be a position-sensitive task. However, contextual information is also crucial to enhance performance even in position-sensitive tasks. In the previous studies [8], [9], [10], they employed contextual

information from multiple sub-regions of feature maps through pyramid pooling or spatial attention. Inspired by [8], we propose the FPF module. Since the pyramid pooling [8] employs the intermediate feature map of the encoder model, such as the existing ResNet, it was designed to employ multiple pooling operations with different kernel sizes. However, LPONet is an encoder-decoder architecture, and there already exist feature maps with multiple resolutions. Particularly, feature maps in the decoder include denser information since it is augmented by a lateral skip connection.

As shown in Figure 2, the proposed FPF module employs m feature maps of the encoder and decoder. To maintain the weight of the decoder's output, the number of channels of each feature map is reduced to C/m using a 1×1 convolutional layer and upsampled to 64×64 resolution. Then, m feature maps are concatenated with the original decoder's output. Finally, we obtain a fused feature map with $64 \times 64 \times 2C$. The number of inputs to the FPF module can be varied. In the experiment, we compare the performance when $m = 4$ using 4×4, 8×8, 16×16, 32×32 feature maps and $m = 2$ using 4×4, 16×16 feature maps.

C. IOAF

Occlusion is a primary reason for the deterioration of the performance of face recognition or face analysis. The proposed network employs a shared layer, so we can expect that landmark and occlusion tasks have complemented each other. However, a more reliable way is to explicitly use occlusion information. In stacked HG [7], heatmap information is fused when passing the output of i-th HG to $(i+1)$-th HG. Inspired by that, we propose the IOAF module.

The IOAF module generates input of $(i+1)$-th HG using heatmap and occlusion predictions of i-th HG, as shown in Figure 2. Formally, let the input and output of i-th HG be $X_I^{(i)}, X_O^{(i)}$, then our IOAF module can be formulated as Equation 1.

$$\begin{aligned} \mathrm{IOAF}\left(X_I^{(i)}, X_o^{(i)}\right) &= X_I^{(i+1)} \\ &= \left(f_W\left(X_o^{(i)}\right) + f_W\left(\hat{L}_{hm}^{(i)}\right) + X^{(i-1)}\right) \| \hat{O}^{(i)} \end{aligned} \tag{1}$$

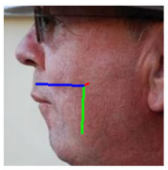
where f_W is 1×1 convolution, $\hat{L}_{hm}^{(i)} = f_W\left(X_O^{(i)}\right)$, $\hat{O}^{(i)} = f_W\left(X_O^{(i)}\right)$. The IOAF module contributes to the overall performance of LPONet by explicitly employing occlusion prediction.

D. GIMBAL-LOCK-FREE HEAD POSE

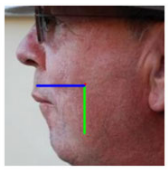
Euler angles representation has the advantage of being intuitive and easy to understand. However, it has a gimbal-lock issue, in which there are numerous rotation parameters for the same visual head pose appearance [11]. Figure 3 illustrates the difference between the ground-truth (GT) label and predicted values in terms of Euler angles and rotation matrix. In the rotation matrix visualized with three axes, two values are almost the same, while in Euler angles, the difference is 28° on average. This ambiguity can have a negative effect on optimizing networks [11].

We use a rotation matrix, instead of Euler angles to address the problem. Previous studies [11], [12] proposed a continuous conversion function between 6D representation and rotation matrix based on the Gram–Schmidt-like process in Equation 2. It converts 6D representation to the 3×3 rotation matrix, satisfying the orthogonality constraint. We modify the head pose head of the LPONet to output six parameters and obtain rotation matrix $R \in \hat{0}\mathrm{R}^{3\times3}$ from the six parameters by applying Equation 2. In the experiment, we compare the performances of these two representations.

$$\begin{aligned} f_{GS}\left(\begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix}\right) &= \begin{bmatrix} | & | & | \\ b_1 & b_2 & b_3 \\ | & | & | \end{bmatrix} = R \\ b_1 &= \frac{a_1}{\|a_1\|}, \\ b_2 &= \frac{u_2}{\|u_2\|}, u_2 = a_2 - (b_1 \cdot a_2)\, b_1, \\ b_3 &= b_1 \times b_2 \end{aligned} \tag{2}$$



(a) GT



(b) Predicted

	Pitch	Yaw	Roll
GT	24	84	20
Pred	-17	89	-18
Error	**41**	**5**	**38**

(c) Error in Euler angles(°)

FIGURE 3. Example about ambiguity in Euler angles. In (a), (b), we plot head pose axes with rotation matrix.

E. LOSS FUNCTIONS

We define loss functions for each task to train the LPONet. Head pose loss is defined as a mean absolute error (MAE) like Equation 3.

$$\mathcal{L}_h(\widehat{P}) = \frac{1}{N}\sum_{i=1}^{N}\left|P_i - \hat{P}_i\right| \tag{3}$$

where N is the mini-batch size, P_i is a GT head pose, $\hat{P}_i$ is a predicted head pose. P_i can have 3 or 9 parameters depending on the head pose representation. In the case of a rotation matrix, the following geodesic loss [11] can be used.

$$\mathcal{L}_g(\widehat{R}) = \frac{1}{N}\sum_{i=1}^{N}\cos^{-1}\left(\frac{\mathrm{tr}\left(R_i\hat{R}_i^T\right) - 1}{2}\right) \tag{4}$$

where R_i represents a GT rotation matrix and $\hat{R}_i$ represents a predicted rotation matrix. The geodesic loss represents the geodesic distance between the rotation matrix and is in $[0, \pi]$.

We followed the heatmap method for the landmark detection task. We converted 68 landmark coordinates to $64 \times 64 \times 68$ landmark heatmap as a GT label to generate a

heatmap. Each pixel of the heatmap is set to 0 as a background pixel or 1 as a foreground pixel depending on the existence of the landmark point. To smooth the heatmap, 7×7 gaussian kernel is applied channel-wise. Based on this heatmap, landmark heatmap loss is defined as a pixel-wise mean squared error (MSE) like the following equation.

$$\mathcal{L}_l(\hat{L}) = \frac{1}{NHWC} \sum_{i=1}^{N} \sum_{j=1}^{HWC} \left(\omega L_{i,j} - \widehat{L}_{i,j}\right)^2 \tag{5}$$

where H, W, and C represent the height, width, and channels of a heatmap, respectively. L_i represents a GT landmark heatmap and $\hat{L}_i$ represents a predicted landmark heatmap. We also applied a weighted loss map [28] to focus on foreground pixels and difficult background pixels which is close to foreground pixels. First, the loss map mask is defined as Equation 6.

$$M = \begin{cases} 1 & \text{where } L_d \geq 0.2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where L_d is generated from GT heatmap L by 3×3 dilation. The loss map mask M assigns foreground pixels and difficult background pixels 1, and other pixels 0. Then, the weighted loss map is formulated by substituting the squared error term in Equation 5 to the weighted squared error like Equation 7.

$$WSE = \left(\omega L_{i,j} - \widehat{L}_{i,j}\right)^2 \otimes (W \cdot M + 1) \tag{7}$$

where $\otimes$ is element-wise multiplication, and W is a scalar hyperparameter. We set W to 10 in our experiments.

The pixel-wise MSE loss has an issue in that if the error is reduced below a certain level, the gradients become rapidly smaller and the training is converged early. Previous studies [27], [28] proposed new loss functions based on the logarithm function to address the problem. We suggest another solution, MSE loss with heatmap scaling. The naïve method to increase gradients is to scale loss value directly. However, it is easy to fall into the local minima in the early stage since it increases both the magnitude of the error and the slope of the gradient function. However, heatmap scaling multiplies GT heatmap by ω to extend the domain range of loss function from [0, 1] to [0, ω]. This prevents the error from rapidly decreasing while maintaining the slope of the gradient function. We plot the training losses and test errors to compare the loss scaling and heatmap scaling methods, as shown in Figure 4. In the case of the loss scaling, it shows a significant loss in the early stage and rapidly converges. Therefore, the final error is the largest. It can be interpreted as falling into the local minima. In the case of the heatmap scaling, it shows a stable loss curve and the smallest error.

Occlusion segmentation loss is defined as a pixel-wise binary cross entropy loss like Equation 8.

$$\mathcal{L}_o(\widehat{O}) = -\frac{1}{NHW} \sum_{i=1}^{N} \sum_{j=1}^{HW2} O_{i,j} \log \hat{O}_{i,j} \tag{8}$$

where O_i represents a GT occlusion mask, and $\hat{O}_i$ represents a predicted occlusion mask.

Finally, our multitask loss is defined as a weighted sum of three task-specific losses like Equation 9.

$$L(\widehat{L}, \hat{P}, \widehat{O}) = \alpha \mathcal{L}_l(\widehat{L}) + \beta \mathcal{L}_h(\widehat{P}) + \gamma \mathcal{L}_o(\widehat{O}) \tag{9}$$

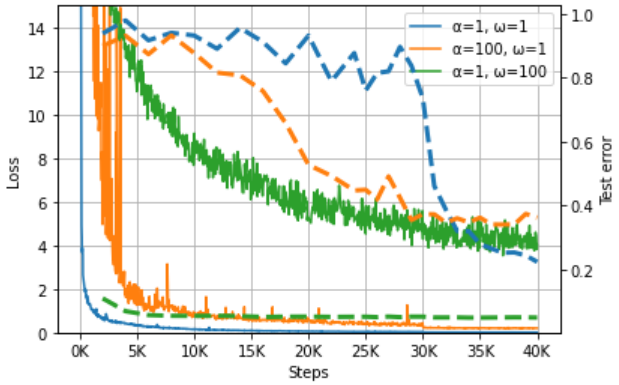


FIGURE 4. Training curves of landmark loss (solid) and test error (dotted). Blue is baseline, orange is loss scaling, and green is heatmap scaling.

IV. EXPERIMENTS

In this section, we describe the experimental environment and evaluation findings of the proposed LPONet.

A. DATASETS

No public dataset that provides a landmark, head pose, and occlusion labels, simultaneously. Therefore, we use two datasets; 300W-LP and C-CM. 300W-LP provides 61,225 yaw-augmented face images and corresponding 68 landmark and 3D head pose labels. C-CM is constructed from CelebAMask-HQ, which provides 30,000 high-resolution images and annotated segmentation masks of facial parts. C-CM is developed by Voo et al. [14] who converted the original masks to binary masks with only face and background classes. C-CM does not provide landmark or bounding box labels.

Two test datasets per task are used for evaluation. (300W, WFLW), (AFLW2000, BIWI), and (COFW, RealOcc-Wild) are employed for the landmark, head pose, and occlusion tasks, respectively. 300W annotates five datasets, including LFPW, AFW, HELEN, XM2VTS, and IBUG, with 68 landmarks. We follow [17], which divides the dataset into a training set with 3148 images and a test set with 689 images. The test set is split into 554 and 135 images as a common and challenging set, respectively. WFLW, which was introduced by Wu et al. [17], provides bounding boxes and 98 landmarks of 10,000 faces in the WIDER FACE dataset [56]. It consists of training and test set, each has 7500 and 2500 faces. The test set is split into six subsets; pose, expression, illumination, make-up, occlusion, and blur. Our model targeted 68 landmarks, so we converted 98 landmark points to 68.

AFLW2000 annotates the first 2000 images from AFLW [57] dataset with 3D head pose and landmark. We excluded 31 images whose poses are not in [−99, 99] degrees. BIWI annotates 24 videos, 15,678 frames with a head pose, collected from 20 subjects in a controlled environment. It provides a head pose label as a rotation matrix.

COFW annotates 1345 occluded faces with 29 landmarks and an occlusion segmentation mask. Currently, only a training set with 500 images is publicly available, as mentioned in [14]. Its definition of occlusion is slightly different from that of the C-CM we follow. Thus, we modified some segmentation masks. The beard was modified from the background to the face and the transparent lens of glasses was modified from the face to the background. RealOcc-Wild was introduced by Voo et al. [14]. It consists of 270 high-resolution occluded face images and segmentation masks.

We cropped the faces from the images with bounding boxes or landmark labels. If those labels are not provided (e.g., C-CM, BIWI, RealOcc-Wild), we use LSFD [58] detector and get face bounding boxes. We also enlarged the bounding boxes by 10% on both sides.

Furthermore, Masked 300W and AFLW2000-SO datasets are used to assess the occlusion robustness of landmark detection and head pose estimation tasks, respectively. Masked 300W is a synthesized masked face dataset based on a 300W test set. AFLW2000-SO is generated by occluding cropped faces with five types of rectangles (left, top, right, bottom, and middle), inspired by Borghi et al. [21].

B. IMPLEMENTATION DETAILS

The proposed network employs a 256 × 256 input image. We optimized the network with RMSPROP with an initial learning rate of 10^{-4}during training. We iterate for 40K with batch size 64, and the learning rate is decreased to 10^{-5} after 30K iterations. We used a 2-stage training method for fine-tuning the network. In the first stage, we train the network employing both 300W-LP and C-CM datasets. The mini-batch is sampled from each dataset with a 1:1 ratio. In the second stage, the 300W train set or WFLW train set depending on the test set for the landmark, 300W-LP for the head pose, and C-CM for occlusion are employed for training. The mini-batch is sampled from each dataset with a 4:3:3 ratio. Loss weights were experimentally determined while heatmap scaling ($\omega = 60$) is applied. In the first stage, $\alpha = 1$, $\beta = 1$, $\gamma = 10$ are set, and $\alpha = 0.1$, $\beta = 0.1$, and $\gamma = 40$ are set in the second stage. The weighted loss map is only applied in the second stage.

Data augmentation is performed using random rotation (±30°), random translation (±25 pixels), random scale (±10%), horizontal flipping, gaussian blur. We also synthesize occluded faces with masks, hands, and sunglasses. We used the open-source tool MaskTheFace [59] for masks and sunglasses, which uses some facial key points. Because C-CM does not provide landmark labels, we used the landmarks one of our pretrained models predicted.

During inference on landmark coordinates, we use a similar approach in [7] and [28], which is a weighted sum of two locations with the highest scores. The only difference is that we employ heatmap scores as weights, instead of fixed weights.

C. EVALUATION METRICS

To evaluate the landmark detection error, we employed the normalized mean error (NME) metric,

$$NME = \frac{100}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\frac{\left\|\left(\hat{l}_{i,j} - l_{i,j}\right)\right\|_2}{d} \tag{10}$$

where N represents the number of images and $\hat{l}_{i,j}$, $l_{i,j}$ represent the j-th predicted and GT landmark coordinates. The error is normalized with the inter-ocular distance, d.

To evaluate the head pose estimation error, we used the MAE metric,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{p}_i - p_i\right\| \tag{11}$$

where $\hat{p}_i$, p_i represent the predicted and GT head pose parameters.

To evaluate the occlusion segmentation accuracy, we used the mean intersection of union (mIoU) metric,

$$mIoU = \frac{1}{2N}\sum_{i=1}^{N}\sum_{c=1}^{2}\frac{\left(\hat{o}_{i,c} \cap o_{i,c}\right)}{\left(\hat{o}_{i,c} \cup o_{i,c}\right)} \tag{12}$$

where $\hat{o}_{i,c}$ and $o_{i,c}$ represent the predicted and GT segmentation masks for class c.

D. ABLATION STUDY

In this subsection, we conduct the ablative study for the three proposed auxiliary components of the network; FPF module, IOAF module, and gimbal-lock-free head pose.

1) FPF

The goal of the FPF module is to advance the performance of the decoder by fusing feature maps with different receptive fields. We designed two variants; FPF_Full, which uses all feature maps ($m = 4$) in the decoder, and FPF_Half, which uses two feature maps ($m = 2$) in the decoder. The landmark and occlusion accuracy were increased in both cases compared to the baseline, as shown in Table 1. Particularly, occlusion segmentation accuracy is significantly increased in the RealOcc-Wild dataset. FPF_Half achieved around 1% higher accuracy than the baseline. Meanwhile, the FPF module also affects the head pose accuracy. FPF_Half shows almost the same accuracy as the baseline, but FPF_Full lowers the accuracy.

TABLE 1. Ablative results of the LPONet with regard to the FPF and IOAF modules.

Method	Landmark NME (%)		Head Pose MAE (°)		Occlusion mIoU (%)		TME
	300W test	WFLW test	AFLW2K	BIWI	COFW train	RealOcc-W	
LPONet-Baseline	4.76	4.28	3.49	4.31	93.06	92.37	5.24
LPONet + FPF_Full	4.74	4.25	3.57	4.42	**93.23**	93.10	5.11
LPONet + FPF_Half	4.74	4.24	**3.46**	4.29	93.14	**93.36**	5.04
LPONet + FPF_Full + IOAF	**4.72**	4.19	3.53	4.38	93.17	93.28	5.06
LPONet + FPF_Half + IOAF	**4.72**	**4.18**	3.56	**4.22**	93.17	93.31	**5.03**

To compare overall performances, we also defined total mean error (TME) as Equation 13, and the FPF_Half module shows better TME than the FPF_Full.

$$TME = (NME + MAE + (100 - mIoU))/3 \tag{13}$$

To sum up, the FPF module makes a positive impact on the landmark and occlusion tasks, and it is more effective in the occlusion task. Additionally, the FPF_Half module can achieve better performance than the FPF_Full in terms of both accuracy and efficiency.

TABLE 2. Error comparison of two head pose representations in AFLW2000 and BIWI datasets.

Pose Rep.	MAE in Euler (°)			Geo. Dist. (rad)		
	AFLW 2K	BIWI	Mean	AFLW 2K	BIWI	Mean
3D Euler	3.52	4.24	**3.88**	0.093	0.142	**0.118**
6DRotRep	3.49	4.31	3.90	0.093	0.146	0.120

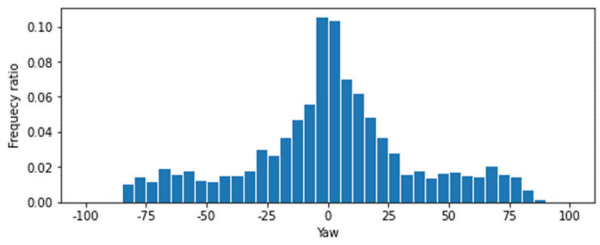


FIGURE 5. Yaw angle distribution in AFLW2000.

2) IOAF

We add IOAF modules to the networks; LPONet + FPF_Full and LPONet + FPF_Half. For both cases, it indicates an improvement in accuracy for all three tasks. Although the improvement is marginal in terms of TME, it is meaningful that IOAF increases the accuracy in a simple way. That means concatenating the occlusion mask to the input of the HG subnetwork in LPONet could help increase overall performance.

3) GIMBAL-LOCK-FREE HEAD POSE

Theoretically, 3D Euler angles have a gimbal-lock problem, as we mentioned in Section III-D. In [11], head pose estimation with 6D rotation representation achieved better accuracy than that with Euler angles. However, there was little difference in accuracy between those representations in our experiment using LPONet-Baseline, as shown in Table 2. We also employed the geodesic distance introduced in [12] as another evaluation metric. Even in this case, the difference between the two representations was negligible.

The rotation matrix has merit to address the ambiguity of Euler angles and supports the full range of head pose. However, we could not determine the merit since the test sets rarely include the faces whose poses are out of the range of ±90°, as shown in Figure 5. Therefore, we can conclude that if the head pose is in the range of ±90°, both representations have no meaningful difference in terms of accuracy.

E. COMPARISON WITH SOTA

The proposed LPONet is the first multitask model for FaceLPO tasks. Thus, it is compared with SOTA for each task since there is no comparison group for the same method. We selected the LPONet with FPF_Half and IOAF modules as our best model because it showed the best accuracy in the ablative study. LPONet only achieved the best accuracy in the AFLW2000 head pose dataset and a little bit lower accuracy in other datasets, as shown in Table 3. In the following subsections, we examine the performance for each task, in more detail.

1) LANDMARK DETECTION

We assessed the landmark NMEs on 300W and WFLW test sets. Table 5 shows the evaluation results of the 300W test set. The proposed LPONet achieved 3.39% NME, which is around 15% lower than SOTA [31]. In Table 4, the proposed model achieved 4.18% NME on WFLW test set, which is the second-best accuracy compared to SOTA methods. It also demonstrated good accuracy across all six subsets. The WFLW test set is more difficult than the 300W dataset according to [28]. Nevertheless, our model showed relatively better results on WFLW than 300W.

2) HEAD POSE ESTIMATION

We evaluated the head pose MAEs on AFLW2000 and BIWI datasets. The evaluated model is the same as the one employed in the landmark evaluation. The proposed LPONet

TABLE 3. Overall comparison of landmark NMEs, head pose MAEs, and occlusion mIoUs with task-specific SOTAs.

Method	Model	#Params (M)	Landmark NME (%)		Head pose MAE (°)		Occlusion mIoU (%)	
			300W test	WFLW test	AFLW2K	BIWI	COFW train	RealOcc-W
SubPixel [31]	2-HG	6.3	**2.94**	**3.72**	-	-	-	-
MNN [6]	Mod. U-Net	-	-	-	3.83	**3.66**	-	-
Voo et al.[14]	SegFormer	84.7	-	-	-	-	**94.87**	**95.16**
LPONet f.t. with 300W	2-HG	6.8	3.39	7.63	3.74	4.32	93.21	93.17
LPONet f.t. with WFLW	2-HG	6.8	4.72	4.18	**3.56**	4.22	93.17	93.31

TABLE 4. Comparison of Landmark detection NMEs(%) on WFLW dataset.

Method	Test	Pose	Expression	Illumination	Make-up	Occlusion	Blur
LAB [17]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
Wing [27]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
HR-Net [48]	4.60	7.94	4.85	4.55	4.29	5.44	5.42
Awing [28]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
LUVLi [60]	4.37	-	-	-	-	-	-
SubPixel [31]	**3.72**	-	-	-	-	-	-
LPONet (Ours)	4.18	**7.05**	**4.44**	**4.23**	**3.95**	**5.16**	**4.94**

TABLE 5. Comparison of landmark detection NMEs(%) on 300W dataset.

Method	Common	Challenge	Full
LAB [17]	2.98	5.19	3.49
DU-Net [61]	2.90	5.15	3.35
HR-Net [48]	2.87	5.15	3.32
Awing [28]	2.72	4.52	3.07
LUVLi [60]	2.76	5.16	3.23
SubPixel [31]	**2.61**	**4.13**	**2.94**
LPONet (Ours)	2.96	5.19	3.39

TABLE 6. Comparison of occlusion segmentation mIoUs(%) on COFW and RealOcc-Wild datasets.

Method	Backbone	#Params (M)	COFW train	RealOcc-Wild
Voo *et al.* [14]	PSPNet	68.1	91.82	91.33
	DeepLabv3+	62.7	92.77	91.01
	SegFormer	84.7	**94.87**	**95.16**
LPONet (Ours)	2-HG	6.8	93.17	93.31

achieved 3.56° MAE on AFLW2000, which outperforms the previous SOTA [6], as shown in Table 7. Particularly, the MAE of the yaw angle was significantly improved by around 15%. However, the MAE on the BIWI dataset is 4.22°, which is 22% lower than the previous SOTA [11]. Even though it is the same model, it demonstrated relatively higher performance on AFLW2000 than on BIWI. It is probably interpreted that the training dataset, 300W-LP, has been biased against AFLW2000 since those two datasets are labeled by the same algorithm [13].

3) OCCLUSION SEGMENTATION

We evaluated the occlusion segmentation mIoUs on COFW and RealOcc-Wild datasets. The evaluated model is the same as the one used in the landmark evaluation. The previous SOTA on these two datasets is only one [14]. Voo et al. [14] benchmarked the performances with PSPNet [8], DeepLapv3+ [63], and SegFormer [49] trained with the C-CM dataset. The proposed LPONet had the second-best accuracy on both datasets in Table 6. Although our model has around 2% lower accuracy than SegFormer, it has an advantage in terms of efficiency because its model size is about 12 times smaller.

F. OCCLUSION ROBUSTNESS

The proposed LPONet can be robust to occlusion because it is based on a multitask learning method that shares layers for all tasks, including occlusion segmentation. To assess the robustness of our model in terms of landmark detection and head pose estimation, we use Masked 300W and AFLW2000-SO datasets.

As shown in Table 8, we achieved the best accuracy on Masked 300W. Compared with the GlomFace [65], our LPONet shows a 23% improvement in performance. It is noted that we augmented 300W training set by synthesizing occluded faces with masks, hands, and sunglasses. According to Zhu et al. [18], they trained their networks without using

TABLE 7. Comparison of head pose estimation MAEs(°) on AFLW2000 and BIWI datasets.

Method	AFLW2000				BIWI			
	Yaw	Pitch	Roll	Mean	Yaw	Pitch	Roll	Mean
HopeNet($\alpha = 1$) [36]	6.92	6.64	5.67	6.41	4.81	6.61	3.27	4.90
FSA-Net [33]	4.50	6.08	4.64	5.07	4.27	4.96	2.76	4.00
QuatNet [62]	3.97	5.62	3.92	4.50	**2.94**	5.49	4.01	4.15
FDN [38]	3.78	5.61	3.88	4.42	4.52	4.70	2.56	3.93
MNN [6]	3.34	4.69	3.48	3.83	3.98	4.61	**2.39**	3.66
6DRepNet [11]	3.63	4.91	3.37	3.97	3.24	**4.48**	2.68	**3.47**
LPONet (Ours)	**2.83**	**4.56**	**3.30**	**3.56**	4.94	4.87	2.85	4.22

TABLE 8. Comparison of landmark detection NMEs(%) on Masked 300W dataset.

Method	Common	Challenge	Full
HGs [7]	8.17	13.52	9.22
FAN [15]	7.36	10.81	8.02
LAB [17]	6.07	9.59	6.76
SRN [64]	5.78	9.28	6.46
SAAT [18]	5.42	11.36	6.58
GlomFace [65]	5.29	8.81	5.98
LPONet-RP	4.74	7.97	5.37
LPONet (Ours)	**4.07**	**6.82**	**4.61**

TABLE 9. Comparison of head pose estimation MAEs(°) on AFLW2000-SO dataset. 'None' refers to the head pose estimation performance on AFLW2000 dataset.

Occluded part	Yaw	Pitch	Roll	Mean
None	2.83	4.56	3.30	3.56
Left	4.47	5.01	3.75	4.41
Top	3.67	5.87	4.39	4.65
Right	5.32	6.13	4.67	5.37
Bottom	2.76	5.00	3.76	3.84
Middle	2.96	4.82	3.42	3.73

masked faces, but using handcrafted occlusion patch. For fair comparison, we trained our network using 300W training dataset augmented by random patch (RP). This network is named LPONet-RP and shows 13% higher accuracy than GlomFace. In consequence, the proposed LPONet shows best performance in occlusion-robust landmark detection task.

Table 9 shows the evaluation results of LPONet using AFLW2000-SO dataset. The dataset has been generated by removing parts of a face image to simulate occlusions as depicted in Figure 6. Because this dataset is first introduced in this work, we could not compare the results with previous SOTAs. Instead, we could provide our results as a baseline for future works. The lowest accuracy was 5.37° MAE when occluded part is right. It is 51% lower performance compared to the original AFLW2000 dataset, which has not been applied simulated occlusions. The occlusion types in the AFLW2000-SO dataset have different impacts on the performance. The left, top, and right part occlusions significantly increase the errors. Specifically, the absence of the left or right part of the face increases the error of the yaw angle, while the absence of the top or right part increases the error of the pitch angle. Alternatively, the bottom and middle part occlusions had a lesser impact on performance than the other three parts.

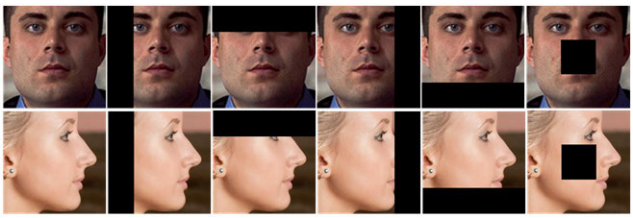


FIGURE 6. Visual examples of AFLW2000-SO dataset. None, left, top, right, bottom, and middle part occlusions from left column.

V. CONCLUSION

In this study, we proposed LPONet, a multitask encoder-decoder network, to examine FaceLPO. It was constructed on a 2-stacked HG network for three tasks. First, the landmark detection task follows a heatmap-based method that regresses a heatmap generated from landmark coordinates. Second, the head pose estimation task is based on direct regression to 3D Euler angles. Third, an occlusion segmentation task is a pixel-wise binary classification with face and background classes. Particularly, we proposed a landmark heatmap scaling approach and experimentally demonstrated that it can help avoid local minima.

Furthermore, we designed three auxiliary components to enhance the LPONet's accuracy. First, we designed an FPF module for the decoder part of our network to employ more contextual information. It demonstrated a performance improvement in landmark and occlusion tasks, which are related to the outputs of the decoder. Particularly, it was

remarkable for the occlusion segmentation task. Second, we proposed an IOAF module that explicitly delivers an intermediate occlusion mask between subnetworks. It also demonstrated an additional improvement in performance. Third, we used a 6D rotation representation for the head pose task instead of 3D Euler angles. Theoretically, Euler angles have intrinsic ambiguity in a large pose. Thus, neural networks have difficulties optimizing accurate head pose. However, in our experiments, there were few differences in the accuracy between Euler angles and 6D rotation representation.

Datasets are crucial for neural network training. Currently, no publicly available dataset supports FaceLPO-related labels simultaneously. Therefore, we trained our network using 300W-LP and C-CM datasets. Additionally, we fine-tuned the network using another landmark dataset, 300W or WFLW, instead of using 300W-LP. In the experiments, we assessed our network, LPONet, for each task. For the head pose, LPONet achieved 3.56° MAE, which is the best accuracy in the AFLW2000 dataset. For landmark, it achieved 4.18% NME, which is the second-best accuracy in the WFLW dataset. For occlusion, it also achieved the second-best accuracy in COFW and RealOcc-Wild datasets.

Furthermore, we showed our network is robust to occlusion in landmark detection and head pose estimation, using Masked 300W and AFLW2000-SO datasets. Although the proposed network did not have the best performances across all datasets, the findings could be satisfactory for the first multitask approach to FaceLPO tasks. Additionally, it is a single network with only 6.8M parameters, making it practical for real-world applications where computational resources are limited.

REFERENCES

[1] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4295–4304.

[2] Y. Wang, H. Yu, J. Dong, B. Stevens, and H. Liu, "Facial expression-aware face frontalization," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 375–388.

[3] S. Kang, J. Lee, K. Bong, C. Kim, Y. Kim, and H.-J. Yoo, "Low-power scalable 3-D face frontalization processor for CNN-based face recognition in mobile devices," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 4, pp. 873–883, Dec. 2018.

[4] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential Siamese network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 773–782.

[5] W. Wan and J. Chen, "Occlusion robust face recognition based on mask learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3795–3799.

[6] R. Valle, J. M. Buenaposada, and L. Baumela, "Multi-task head pose estimation in-the-wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2874–2881, Aug. 2021.

[7] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 483–499.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6230–6239.

[9] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Oct. 2018, pp. 270–286.

[10] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, Korea (South), Oct. 2019, pp. 593–602.

[11] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6D rotation representation for unconstrained head pose estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Bordeaux, France, Oct. 2022, pp. 2496–2500.

[12] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5738–5746.

[13] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," 2015, *arXiv:1511.07212*.

[14] K. T. R. Voo, L. Jiang, and C. C. Loy, "Delving into high-quality synthetic face occlusion segmentation datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4711–4720.

[15] 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. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1021–1030.

[16] 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, Mar. 2016.

[17] 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 Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2129–2138.

[18] C. Zhu, X. Li, J. Li, and S. Dai, "Improving robustness of facial landmark detection by defending against adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 11731–11740.

[19] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, 2013.

[20] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.

[21] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 5494–5503.

[22] 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.

[23] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 17–24.

[24] 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.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[26] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "PFLD: A practical facial landmark detector," 2019, *arXiv:1902.10859*.

[27] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2235–2245.

[28] 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)*, Seoul, Korea (South), Oct. 2019, pp. 6970–6980.

[29] C.-F. Hsu, C.-C. Lin, T.-Y. Hung, C.-L. Lei, and K.-T. Chen, "A detailed look at CNN-based approaches in facial landmark detection," 2020, *arXiv:2005.08649*.

[30] H. Jin, S. Liao, and L. Shao, "Pixel-in-Pixel Net: Towards efficient facial landmark detection in the wild," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3174–3194, Dec. 2021.

[31] A. Bulat, E. Sanchez, and G. Tzimiropoulos, "Subpixel heatmap regression for facial landmark localization," in *Proc. BMVC*, Nov. 2021, pp. 1–15.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," 2018, *arXiv:1801.04381*.

[33] T. Yang, Y. Chen, Y. Lin, and Y. Chuang, "FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1087–1096.

[34] D. F. Dementhon and L. S. Davis, "Model-based object pose in 25 lines of code," *Int. J. Comput. Vis.*, vol. 15, no. 1, pp. 123–141, Jun. 1995.

[35] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3D head pose estimation with convolutional neural network trained on synthetic images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 1289–1293.

[36] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2187–2196.

[37] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," in *Proc. BMVC*, 2020, pp. 1–13.

[38] H. Zhang, M. Wang, Y. Liu, and Y. Yuan, "FDN: Feature decoupling network for head pose estimation," in *Proc. AAAI Conf.*, Apr. 2020, pp. 12789–12796.

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

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1–9.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14.

[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[43] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2017.

[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803.

[45] 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.

[46] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 472–487.

[47] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[48] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, and W. Liu, "Deep high-resolution representation learning for visual recognition," in *Proc. Int. Conf. CVPR*, Jun. 2019, pp. 5686–5696.

[49] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NIPS*, Dec. 2021, pp. 1–14.

[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[51] S. Saito, T. Li, and H. Li, "Real-time facial segmentation and performance capture from RGB input," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 244–261.

[52] Y. Nirkin, I. Masi, A. T. Tran, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," 2017, *arXiv:1704.06729*.

[53] X. Yin and L. Chen, "FaceOcc: A diverse, high-quality face occlusion dataset for human face extraction," 2022, *arXiv:2201.08425*.

[54] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5548–5557.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[56] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.

[57] M. Kostinger, 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 (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.

[58] Y. Kim, J.-H. Roh, and S. Kim, "LSFD: Lightweight single stage masked face detector with a CPU real-time speed," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2021, pp. 1818–1822.

[59] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," 2020, *arXiv:2008.11104*.

[60] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng, "LUVLi face alignment: Estimating landmarks location, uncertainty, and visibility likelihood," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 8233–8243.

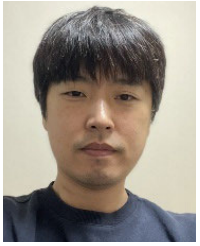
[61] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected U-Nets for efficient landmark localization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 348–364.

[62] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "QuatNet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, Apr. 2019.

[63] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.

[64] C. Zhu, X. Li, J. Li, S. Dai, and W. Tong, "Reasoning structural relation for occlusion-robust facial landmark localization," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108325.

[65] C. Zhu, X. Wan, S. Xie, X. Li, and Y. Gu, "Occlusion-robust face alignment using a viewpoint-invariant hierarchical network architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11102–11111.



YOUNGSAM KIM received the B.S. degree in computer engineering from Chungbuk National University, Cheongju, South Korea, in 2009, and the M.S. degree in information security engineering from the University of Science and Technology, Daejeon, South Korea, in 2011.

From 2010 to 2013, he was a Researcher with the National Institute for Mathematical Sciences (NIMS), Daejeon. He is currently a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon. His research interests include biometrics, machine learning, and context-aware authentication.

JONG-HYUK ROH received the B.S., M.S., and Ph.D. degrees in computer engineering from Inha University, Incheon, South Korea.

He is currently a Principal Researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. His research interests include machine learning, pattern analysis, behavior-based authentication, and computer security.

SOOHYUNG KIM received the B.S. and M.S. degrees in computer science from Yonsei University, Seoul, South Korea, in 1996 and 1998, respectively, and the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2016.

He is currently a Project Leader of the Information Security Research Division, Electronics and Telecommunications Research Institute (ETRI), Daejeon. His research interests include identity management, blockchain security, financial security, and biometrics.