IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 27, 2020, accepted June 12, 2020, date of publication June 23, 2020, date of current version July 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3004359

FLSNet: Robust Facial Landmark Semantic Segmentation

HYUNGJOON KIM, HYEONWOO KIM, JEHYEOK REW, AND EENJUN HWANG, (Member, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Eenjun Hwang (ehwang04@korea.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant 2020R1F1A1074885, and in part by the Ministry of SMEs and Startups (MSS), South Korea, through the Technology Development Program, under Grant S2796678.

ABSTRACT The human face is one of the most viewed visual objects in a person's life and is used for identifying a person through facial landmarks, which includes the eyes, nose, mouth, and ears that make up a face. It is also possible to communicate nonverbally through the movements of facial landmarks; that is, change of facial expression. Thus, facial landmarks play a crucial role in human-related image analysis. Automatic facial landmark detection is a challenging problem in the field of computer vision, and various studies are underway. The emergence of Deep Neural Networks has played an important role in solving difficult problems in computer vision. Semantic segmentation is a field in which images are classified into pixel units and has also developed rapidly by incorporating deep learning. In this paper, we propose a method for accurately extracting facial landmarks using semantic segmentation. First, we introduce a semantic segmentation architecture for sophisticated landmark detection, and datasets composed of facial images and ground truth pairs. Then, we suggest how improve the performance of pixel classification by adjusting the imbalance of the number of pixels according to the face landmark. Through extensive experiments, we evaluated our approach using the metrics pixel accuracy and intersection over union.

INDEX TERMS Facial landmark, semantic segmentation, deep neural networks, network architecture, pixel unbalance, weighted feature map.

I. INTRODUCTION

One of the most common visual objects that people encounter in their lifetimes is human faces. People can identify individuals using facial landmarks: the features that make up a face, such as eyes, nose, and mouth. Humans express their feelings using facial expressions by moving their facial landmarks and can understand the emotions of other people by reading their facial expressions. The appearance and movement of facial landmarks facilitate non-verbal dialogue. Therefore, recognizing facial landmarks plays an essential role in identifying a person or analyzing a person's emotions. Therefore, identifying facial landmarks in images of people is essential for human-related image analysis. Various studies into the detection of facial landmarks have been carried out, but the task is still challenging because of variations in face images caused by factors such as pose, lighting, and occlusions [1].

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

Deep Neural Networks (DNNs) [2] have been shown to be effective in solving challenging computer vision problems such as image classification and object recognition. DNNs have been used in facial landmark detection research, and detection performance has significantly improved with their use [3]. Studies have identified several key points that represent facial landmarks [4]–[7]. These studies generally extract five or 68 landmark points. The five points mark the specific position of major landmarks such as the eyes, nose, and mouth and the 68 points even represent the approximate shape of the face landmarks. These approaches have the advantage of being able to detect landmarks quickly and accurately, and it is even possible to process the data in real-time in a general hardware environment [8], [9]. These technologies can be used for various purposes such as face identification [10], [11] and emotion recognition [12], [13]. Although using five or 68 points is useful and practical, this approach does not detect hair, which plays a significant role in identifying an individual. Furthermore, it does not detect detailed landmark

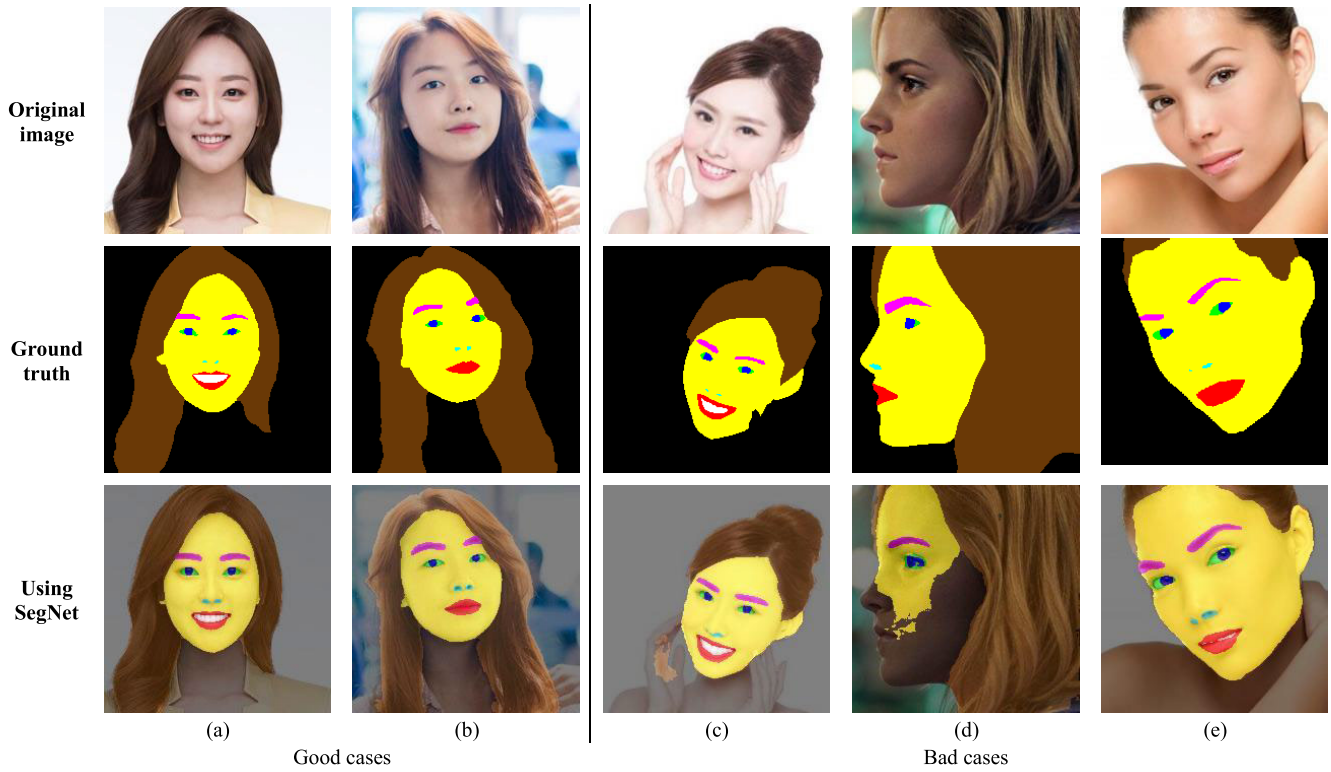


FIGURE 1. SegNet based facial landmark extraction: Bad cases include (c) misunderstanding the background and identifying it as hair, (d) inaccurate detection of side face landmarks, and (e) inaccurate facial landmark detection of the eyebrow and nostril.

shapes such as distinguishing the pupils from the whites of the eyes or detecting the exact shape of the eyebrows.

To detect these more detailed landmarks, in our previous work [14], [15], we used semantic segmentation, a widely-used approach to computer vision, which involves classifying each pixel in an image. In these studies, we used two CNN models; (i) The first model is Faster R-CNN [16] to find a region of the face in the image. The face region is cropped and resized. (ii) The second model is SegNet [17], [18] that performs semantic segmentation using the processed image as input. The model is known to be effective for extracting the detailed shape of an object; its backbone architecture is VGG16 [19]. In terms of pixel accuracy, the model accurately extracts facial landmarks. However, as shown in Figure 1, there are some drawbacks to face landmark detection. For instance, it cannot reliably identify the side face landmarks. Also, if a landmark occupies a small proportion of the pixels in an image, such as a pupil or an eyebrow, the shape may be inaccurate. Such pixels are sometimes identified as belonging to the wrong class.

To overcome these problems and make the model more robust and accurate, we propose a new semantic segmentation architecture, a pixel balancing method, and a pixel classification scheme. We added a shallow semantic segmentation model to refine the results of a basic semantic segmentation model. There is no public facial landmark dataset appropriate for semantic segmentation, so we constructed a dataset manually. The dataset consists of face image and ground truth pairs, with ground truth made up of nine classes that represent specific facial landmarks. The distribution of classes among the pixels is highly unbalanced.

To balance the pixel distribution during the training process, we applied an area-based class weight, which represents the ratio of the number of pixels of a class and the number of occurrences of that class. Finally, we calculated a class weight for pixel-wise classification and applied it to feature maps from the Softmax layer, to improve classification performance. The paper is organized as follows: Section 2 introduces several studies related to facial landmark detection and semantic segmentation. Section 3 describes our algorithm in detail, and in Section 4, we evaluate our approach experimentally. Finally, Section 5 concludes this paper.

II. RELATED WORKS

The purpose of this paper is to describe an approach for accurately extracting facial landmarks using semantic segmentation, regardless of the size of the face or its direction. Facial landmark detection and semantic segmentation are both challenging tasks and have been studied widely.

A. FACIAL LANDMARK DETECTION

Feng and colleagues proposed a framework for detecting facial landmarks in a wild dataset in 2017 [6]. They used two publicly available CNN-based face detectors, dlib and MTCNN, and two proprietary detectors. The bounding boxes established by the face detectors were aggregated to improve detection accuracy. Then, these authors used cascaded shape regressors to estimate pose and preprocess the images.

Finally, the researchers used a cascaded shape regressor to locate the facial landmarks. Ranjan and colleagues described a CNN model called HyperFace in 2017, which performed face detection, landmark localization, pose estimation, and gender recognition [7]. They used one backbone CNN architecture and added various output layers such as regression and classification layers after the fully-connected layer to enable their model to achieve multiple goals. Tang and colleagues detected facial landmarks using semi-supervised learning based on CNNs in 2018 [20]. The model needed only an image as input and solved the problem of occlusion of large areas by detecting visible facial components while existing face detectors failed to detect faces. Zhu and colleagues developed Occlusion-adaptive Deep Networks (ODNs) in 2019 to solve the occlusion problem [21]. ODNs extract feature maps using the residual blocks, and then use the extracted feature maps as input into two CNN modules: A Geometry-aware Module and a Distillation Module. Finally, facial landmarks are extracted from the feature maps concatenated from the modules by passing them through a Low-Rank Learning module and a fully-connected layer. Jackson and colleagues performed semantic segmentation on facial images in 2016 [22]. They designed two semantic segmentation architecture based on fully convolutional networks (FCNs) [23]. The first one is for facial landmark points detection, and the second one is for semantic segmentation. The detected points are used as a guideline for semantic segmentation. Also, there is no pixel annotated facial dataset so that they created pixel annotated dataset using a public facial dataset with 68 key points.

B. SEMANTIC SEGMENTATION

Long and colleagues proposed FCNs, the first deep learning based semantic segmentation model in 2015 [23]. In FCNs, an 11-convolution layer is substituted for the fully-connected layer that was used for image classification in the general CNN model for pixel-wise classification. They also skipped the layer method to improve accuracy during the up-sampling process. To solve the problem of low resolution in bilinear interpolation, which is used in FCNs, Noh and colleagues developed DeconvNet, which added a deconvolutional network symmetrical to the convolutional network [24]. Simultaneously, they used the switch variable concept to remember the location of the max value during the max pooling calculation, and locate its position. Badrinarayanan and colleagues developed SegNet [17], [18], a network that combines the advantages of DeconvNet and U-net. It reduces parameterization by removing the fully-connected layer used in DeconvNet. In addition, it reduces memory cost by using pooling indices as opposed to copying and cropping the entire feature set, as in U-net [25]. Hengshuang Zhao and colleagues proposed Pyramid Scene Parsing Network (PSPNet) in 2017 [26]. In PSPNet, a CNN model is used to obtain the feature map of the last convolutional layer. Then, different sub-region representations are collected from the feature map through Pyramid Pooling Module. The sub-region representations are up-sampled and concatenated with the feature map from the last convolutional layer. Also, they adopted an auxiliary loss to help optimize the learning process. Chen and colleagues proposed a semantic segmentation architecture called DeepLab in 2015 [27]. DeepLab uses ''atrous convolution'' to expand the receptive field without increasing the amount of computation needed, and the use of fully-connected CRFs maximizes the accuracy of pixel-level classification. In 2017, these authors developed DeepLab v2 [28]. The researchers introduced the concept of atrous spatial pyramid pooling (ASPP), in which the last fully-connected layer of the end encoder is replaced with convolution layers with various atrous rates. In DeepLab v3 [29], they used the deeper encoder and proposed a method for obtaining dense feature maps using atrous convolution. Subsequently, they announced DeepLab v3+ [30]. DeepLab v3+ uses the Xception [31] backbone and U-Net architecture. They also introduced the concept of ''atrous separable convolution''. Jun Fu and colleagues proposed the Stacked Deconvolutional Network (SDN) architecture in 2019 [32]. They defined multiple shallow deconvolutional networks as SDN units. SDN units are stacked one by one to integrate contextual information, and to produce fine-grained recovery of localization information. To improve the flow of information and gradient propagation, the researchers designed inter-unit and intra-unit connections.

Jingdong Wang and colleagues proposed a high-resolution network (HRNet) in 2019 [33], [34]. The proposed network generates multi-scale feature maps to strengthen high-resolution representations. Since the last convolution, every different size of feature map is up-sampled to the size of the largest feature map, and all of the same size feature maps are concatenated.

III. PROPOSED METHOD

In this work, we aimed to extract facial landmarks precisely, using semantic segmentation. To achieve this aim, we designed a semantic segmentation architecture and constructed a dataset including face image and ground truth image pairs, for training. Then, we applied an area-based class weights to balance the pixel distribution between classes and introduced weighted feature maps for reflecting the characteristics of the data. The overall flow of our FLSNet model is shown in Figure 2. A detailed description of each step is presented in the next section.

A. ARCHITECTURE

The model architecture described in this paper consists of two semantic segmentation models as shown in Figure 2. The first semantic segmentation model extracts facial landmarks roughly as feature maps, and these feature maps are then used as the inputs to the second model. As shown in Figure 3, feature maps from the encoder in the first model were concatenated with feature maps from the corresponding encoder of the second model, which has the same size feature maps.

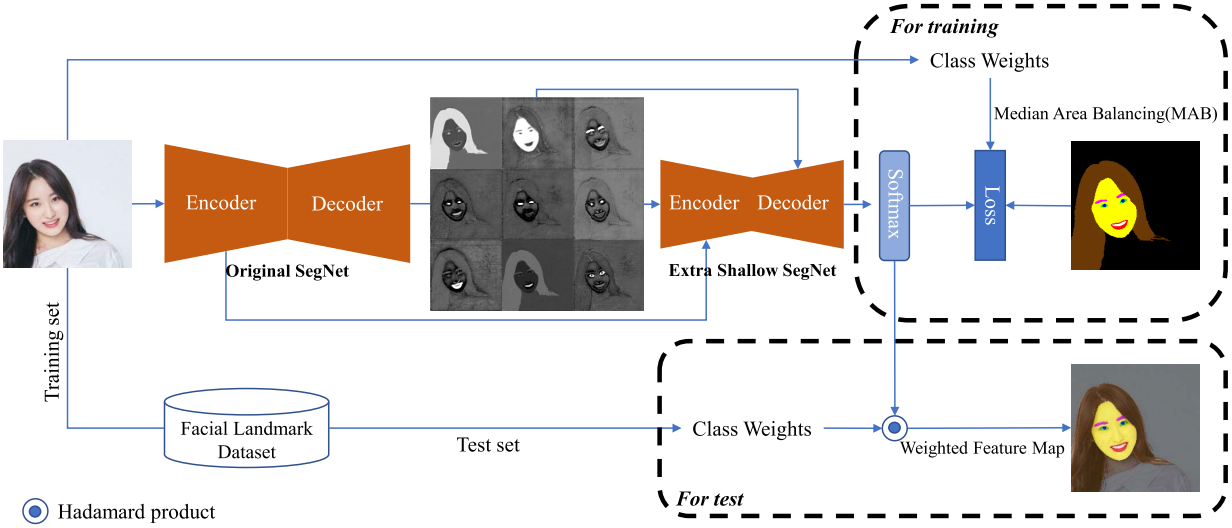


FIGURE 2. Overall architecture of our FLSNet.

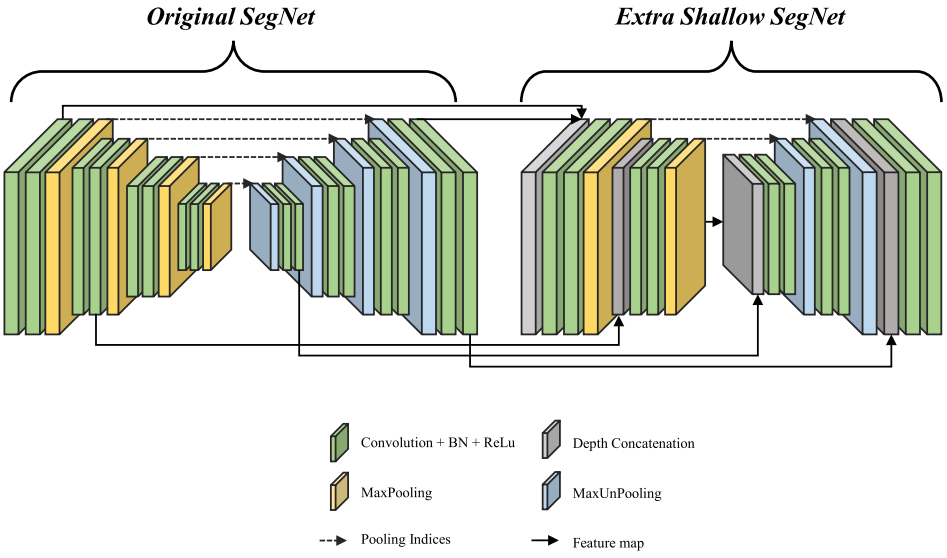


FIGURE 3. Semantic segmentation architecture of our FLSNet.

The feature maps from the decoder are similarly concatenated with corresponding feature maps.

The input of the second model is the feature maps from the decoder of the first model, and the second model decoder obtains feature maps from the encoder of the second model. The feature maps, therefore, reflect the characteristics of adjacent feature maps. Thus, we do not need to connect all the feature maps densely. Therefore, we concatenate only the encoder feature maps to the encoder feature maps, and the decoder feature maps to the decoder feature maps.

The second model is shallower than the first because the final feature maps from the first model extract facial landmarks reasonably well if trained properly. The second model, therefore, does not need to learn all of the features of the input images; it just corrects some minor differences between the results of the first model and ground truth. In Figure 3, the backbone model is SegNet with VGG16. We set the depth of the second model to almost half that of the first model, and the basic architecture is SegNet, as in the first model.

B. DATASET

There are no public data appropriate for the semantic segmentation of facial landmarks, so we manually constructed a facial landmark dataset. The dataset consists of 386 face images from the Figaro hair dataset [35], 323 from the Helen dataset [36], 107 from the AFW dataset [37], 151 from the LFPW dataset [38], 106 from the LFW dataset [39], 124 from the 300W dataset [40] and 592 face images, including various ages, genders, and races. Based on the colors shown in Table 1, we constructed ground truth images corresponding to the face images as shown in Figure 4. Figure 4 (a) shows the original face images, and Figure 4 (b) shows the ground truth of the corresponding face images.

(a) Part of facial image dataset

(b) Ground truth images

FIGURE 4. An example facial landmark dataset: Collected images on the first row, the Figaro hair dataset images on the second row, and the Helen dataset images on the last row.

TABLE 1. R, G, B values of facial landmarks.

	R	G	B	Color
Hair	106	57	6	
Skin	255	255	0	
Eyebrow	255	0	255	
Pupil	0	0	255	
White	0	255	0	
Nostril	0	255	255	
Lip	255	0	0	
Inner mouth	255	255	255	
Background	0	0	0	

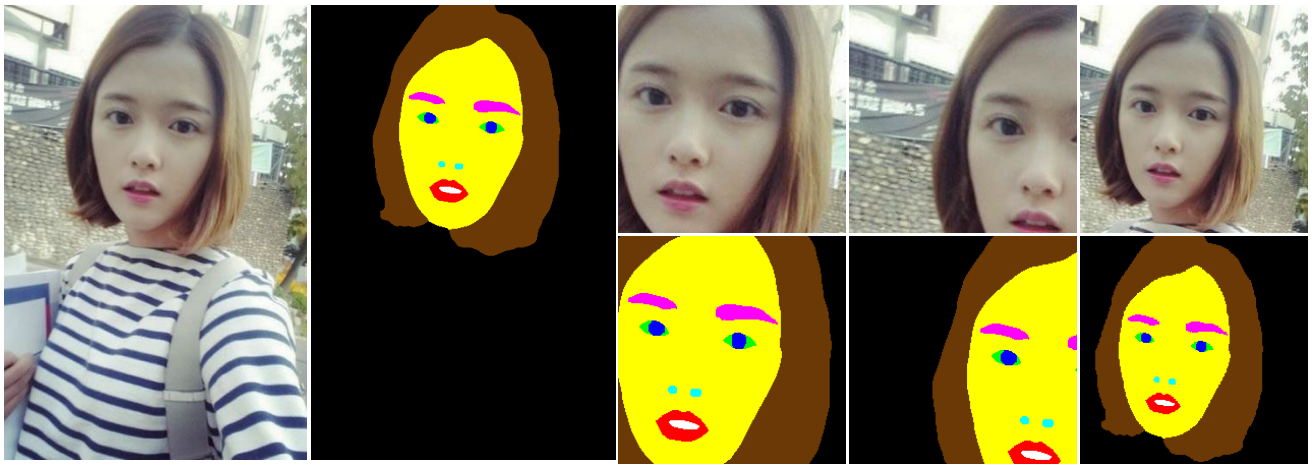
C. TRAINING

Images can have a different sized face if images are taken as full-length photos, or zooms in to the face. Particularly in the case of full-length and high-resolution images, the face can be too small to use directly. Using the face from these images leads to distortion when resizing the images, and some landmarks may be smaller than the convolution filter, making it hard to train the model. To make the model robust to size, direction, and some obscuration of the face, we needed to train our model with images taken at various sizes and directions. To obtain a variety of facial images, we cropped the face images in various sizes and locations. Images cropped at various locations may be missing some landmarks. A model trained with these data could extract facial landmarks well, even when some landmarks were missing as in Figure 5.

As previously mentioned, the model consists of two semantic segmentation models. The second model uses feature maps from the first model. If the performance of the first model is reasonable, it means the model creates fine-grained feature maps. When using high-quality feature maps, the second model will be trained faster. Therefore, we trained Seg-Net with the VGG 16 model first, and then trained a shallow semantic segmentation network using the pre-trained model.

D. METRIC

The loss function used in this work was pixel-wise-cross-entropy, a metric that is widely used for the semantic segmentation task. However, the pixel distribution of the dataset



(a) The original image pair (b) Cropped image pairs

FIGURE 5. Cropped images and corresponding ground truth pair examples. In (b), the first and second images are the same size, but are cropped in different locations, while the last image is of a different size and location.

for facial landmark semantic segmentation is significantly unbalanced. To deal with this imbalance, we used weighted pixel-wise cross-entropy. In previous work [14], [15], we calculated class weights by using median frequency balancing (MFB) method [41]. In the method, the *frequency* of a class c represents the ratio between the number of pixels in the class and the total number of pixels in the dataset images that contain that class. The class weights for N classes are calculated by Equation 1.

$$\begin{aligned} frequency\,(c) &= \frac{NoP(c)}{NoP(I_c)} \\ weight(c) &= \frac{\text{median}(F)}{frequency(c)} \end{aligned} \tag{1}$$

Here, $NoP(c)$ indicates the number of pixels in a class c of N classes, and $NoP(I_c)$ represents the total number of pixels in the set of images that contain the class c. F is a set of N *frequencies*.

Intuitively, if the number of pixels occupied by a specific landmark class is small and the class appears in many images, then, its frequency becomes smaller, and weight becomes larger. Although this approach helps to solve the problem of imbalance among pixels in different classes, when that the class has a very small number of pixels, weight becomes excessively large, even though the class does not occupy much of the image. This situation can lead to overfitting, so it is necessary to limit the weight of the class to prevent it. Thus, when calculating class weight, we consider the ratio between the number of class pixels and the class occurrence; that is, how many times the specific class cluster appears in all images. For example, in the case of a face image of a person looking straight ahead, as shown in the first image in Figure 1 (a), the class occurrence is as follows: one background, one hair, two eyebrows, two whites of eye, two pupils, two nostrils, one lip, and one inner mouth. In the face in the profile shown in the fourth image of Figure 1 (d), there is one background, one hair area, one eyebrow, one white of eye, one pupil, one nostril, one lip and zero inner mouth. The class weight based on the occurrence is calculated using Equation 2.

$$\begin{aligned} area\,(c) &= \frac{NoP(c)}{Occurence(c)} \\ weight(c) &= \frac{\text{median}(A)}{area(c)} \end{aligned} \tag{2}$$

where $Occurrence(c)$ is the number of occurrences of a given class c, A is a set of N *areas*.

As shown in the equation, $area(c)$ means the number of pixels when given class c is once occurred, which means the average area occupied by the class c. MFB gives the large class weight when class c is small. On the other hand, the area-based class weight gives the large class consider not only the size of the class but also its occurrence. Therefore, it prevents a very small class from having a large class weight. We refer to this area-based class weight calculation method as Median Area Balancing (MAB).

Figure 6 shows the comparison of the proposed weight with the median frequency balancing weight in an extremely unbalanced dataset. Figure 6 (a) shows the pixel distribution among the classes, and Figure 6 (b) shows the class weight calculated using two methods. As shown in the figure, the pixel distribution among classes is extremely unbalanced, and the large weights that appear under the median frequency balancing are reduced.

E. CLASSIFICATION METHOD

The pixel classification result of the semantic segmentation is determined from the feature maps that pass through the Softmax layer. The feature maps have the same number of channels as the number of classes, and each channel represents a specific class. If the feature maps computed from the

Softmax layer are called *FM*, the class can be expressed as $\text{Class} = \frac{\text{argmax}}{\text{c}} \{FM\ (\text{c})\}$, where c is the channel.

When classification performance is measured by Intersection over Union (IoU), which is a popular evaluation metric for semantic segmentation, the IoU of the class that occupies large pixel areas in the image has a high value because it can generally be trained easily. In the facial landmark dataset, for instance, the number of pixels of hair, skin, and background classes occupy more than 95% of the image pixels Figure 6 (a). The IoU of these three classes is the highest of all of the classes. The channels representing the corresponding classes amongst the feature maps obtained from the Softmax layer are more accurate than other classes. When classifying boundaries where several classes meet, the values of the channels corresponding to the classes may appear similar. In general, class feature maps with high pixel occupancy are more accurate than those with low pixel occupancy. Therefore, even if the value of the low occupancy class is slightly higher, there is a high probability that it belongs to a high occupancy class. We improved the quality of semantic segmentation by assigning weights in order to select a feature map of a high occupancy class when the values of feature map channels are similar. In Section 3. D, we introduced the area of a landmark class, the average number of pixels of the class in the images. Area is the number of pixels when the class occurs once, and it becomes larger as the pixel occupancy of the class becomes larger. Equation 3 shows how to weight feature maps using area.

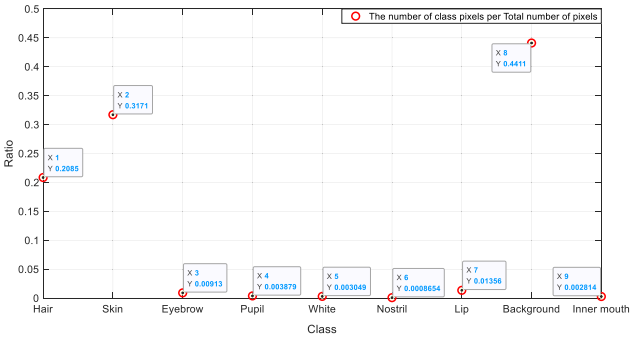
$$FM_{\text{w}}(\text{c}) = \frac{area(\text{c})}{\frac{1}{\text{N}}\sum_{\text{k}=1}^{\text{N}} area(\text{k})} \times FM_{\text{o}}(\text{c}) \tag{3}$$

Here, FM_{o} and FM_{w} denote the original and weighted feature maps, respectively and N denotes the number of landmark classes in the dataset. As shown in Figure 2, the calculated weights are multiplied channel-wise with the feature maps and the result is used for pixel-wise classification.

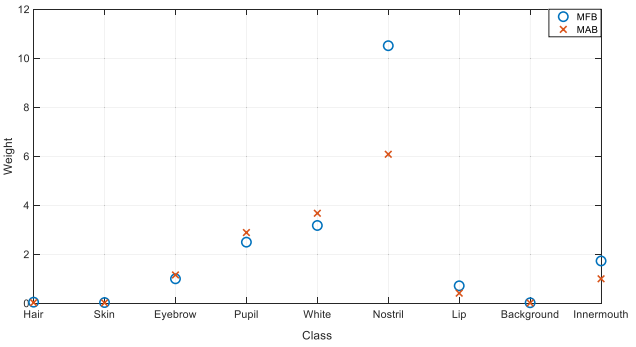
IV. EXPERIMENTS

In the previous section, we described our pixel annotated facial landmark dataset and three methods that we used to build our FSLNet for improving semantic segmentation accuracy, which are class weights, shallow network and weighted feature map. These methods can be used individually or in combination. In this section, we describe several experiments that we performed to evaluate the performance and applicability of our model. At first, we build a variety of combinations of the three methods and compare their performances to show that using them all gives practically the best performance. Then, we compare our method with other well-known semantic segmentation methods. Finally, we show that our model performs well for other datasets.

For quantitative evaluation of landmark extraction, we used two metrics: pixel accuracy and IoU. Pixel accuracy and IoU



(a) The pixel distribution among the classes



(b) Two kinds of class weights

FIGURE 6. Comparison of weights calculated by MFB and MAB.

can be defines as follows:

$$\text{Pixel accuracy} = \text{TP}/(\text{TP} + \text{FN})$$
$$\text{IoU} = \text{TP}/(\text{TP} + \text{FN} + \text{FP})$$

where TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively.

In image segmentation, the IoU is usually preferred over pixel accuracy as it is not as affected by the class imbalances that are inherent in foreground/background segmentation. Overall, our method improves IoU while maintaining the pixel accuracy.

The pixel-annotated facial landmark dataset, as discussed in the Dataset section, was used for network training. The entire dataset consisted of 1,789 facial images. In general, the effectiveness of training increases as the size of the dataset increases. So, to increase the size of the dataset, we first divided it into a training set of 1,432 images and a training set of 357 images. Then, by cropping the images in each set in diverse ways, we increased the number of images in the training and test sets to 59,428 and 5,876, respectively.

The experiments were conducted on an Intel Xeon E5-2680v4 CPU, with 128GB DDR4 memory and two NVIDIA RTX TITAN. The mini-batch size was 48 on VGG16 backbone networks and 12 on other deeper networks. The number of epochs was 500 for all models. When training the shallow network-added model, we set the number of epochs to 100 and used our class weights. The backbone architecture was SegNet with VGG16.

TABLE 2. Pixel accuracy comparison of three methods in facial landmark extraction.

MFB: Median frequency balancing, MAB: Median area balancing

Methods			Classes									
Class Weights	*Shallow Network*	*Weighted Feature Map*	*Hair*	*Skin*	*Eyebrow*	*Pupil*	*White*	*Nostril*	*Lip*	*Background*	*Inner mouth*	*Average*
MFB	X	X	0.967	0.944	0.940	0.960	0.949	0.991	0.926	0.952	0.760	0.932
MFB	X	○	0.926	0.970	0.820	0.920	0.836	0.860	0.832	0.959	0.671	0.866
MFB	○	X	0.985	0.944	0.967	0.966	0.958	0.987	0.959	0.967	0.903	0.961
MFB	○	○	**0.955**	0.965	0.833	0.946	0.871	0.845	0.899	0.959	0.842	0.902
MAB	X	X	0.971	0.943	**0.988**	0.961	0.971	0.991	0.954	0.966	0.876	0.958
MAB	X	○	0.939	0.974	0.922	0.939	0.885	0.827	0.850	0.967	0.760	0.896
MAB	○	X	0.980	0.965	0.988	**0.977**	**0.975**	**0.995**	**0.986**	**0.981**	**0.934**	**0.975**
MAB	○	○	0.958	**0.982**	0.939	0.965	0.915	0.812	0.957	0.981	0.861	0.930

TABLE 3. IoU comparison of three methods in facial landmark extraction.

MFB: Median frequency balancing, MAB: Median area balancing

Methods			Classes									
Class Weights	*Shallow Network*	*Weighted Feature Map*	*Hair*	*Skin*	*Eyebrow*	*Pupil*	*White*	*Nostril*	*Lip*	*Background*	*Inner mouth*	*mIoU*
MFB	X	X	0.883	0.913	0.664	0.822	0.456	0.263	0.822	0.943	0.209	0.664
MFB	X	○	0.845	0.927	0.693	0.870	0.684	0.530	0.794	0.937	0.611	0.766
MFB	○	X	0.910	0.927	0.678	0.823	0.495	0.391	0.841	0.946	0.561	0.730
MFB	○	○	0.867	0.911	0.723	0.872	0.679	0.617	0.835	0.921	0.745	0.797
MAB	X	X	0.876	0.921	0.617	0.811	0.463	0.276	0.855	0.958	0.340	0.680
MAB	X	○	0.870	0.901	0.753	0.879	0.704	0.562	0.823	0.950	0.709	0.795
MAB	○	X	**0.920**	**0.931**	0.712	0.806	0.615	0.443	0.826	0.909	0.712	0.764
MAB	○	○	0.913	0.922	**0.828**	**0.886**	**0.790**	**0.697**	**0.873**	**0.916**	**0.790**	**0.846**

A. EVALUATING THREE METHODS

In the first experiment, we evaluate the effectiveness of the three methods in the extraction of facial landmarks. By applying eight different combinations of three methods to the original SegNet architecture, we built eight different models. We compared their facial landmark extraction performance in terms of pixel accuracy and IoU. The baseline was the original SegNet that did not use any of our methods.

Tables 2 and 3 show their pixel accuracy and IoU for nine different facial landmark classes, respectively. The first and last rows in the tables represent the original SegNet and our FLSNet, respectively. The original SegNet showed very low IoU scores for small-sized landmarks such as eyebrows, whites of eyes, and nostrils, even though their pixel accuracies were higher than 0.9. Figure 1 shows such examples in (c) and (e). In the figure, the nostrils and eyebrows were extracted too larger than the ground truth. On the other hand, our FLSNet gave the best IoU scores for major landmark classes while maintaining the pixel accuracy.

Next, we examine how each method contributed to improving the performance of facial landmark extraction. For this, we calculate the performance ratio when each method is not used and when it is not used. Hence, if the ratio is greater than 1.0, the performance is improved by the method.

1) CLASS WEIGHTS

To consider different sizes of facial landmarks, we proposed a new area-based class weight method called MAB. To see the effect of our class weight method, we calculated the performance ratio when MAB was used and when MFB was used using the tables 2 and 3 assuming the rest of the conditions are the same. Table 4 shows the ratio of pixel accuracy and IoU for each class. As we mentioned earlier, the ratios greater than 1.0 indicate improved performance. The table shows that the pixel accuracy improved slightly, while IoU improved significantly especially for small landmarks.

Figure 7 shows two landmark extraction results by the original SegNet trained with MFB and our MAB for one face.

In the figure, while the nostril in (c) is clearly larger than the ground truth in (b), the nostril in (d) is very similar to the ground truth. Hence, the IoU of the original SegNet is significantly lower than that of SegNet with MAB. Overall, we achieved more accurate facial landmark extraction, in particular for small landmarks, by using our class weight method.

2) ADDITONAL SHALLOW NETWORK

The purpose of adding a shallow network is to fine-tune the feature maps from the first semantic segmentation network. The benefits of using the shallow network can be seen in Table 5. In the table, pixel accuracy improves less than 5%,

TABLE 4. Landmark extraction performance comparison of class weight method.

MFB: Median frequency balancing, MAB: Median area balancing

Metrics	Methods		Classes									
	Shallow Network	*Weighted Feature Map*	*Hair*	*Skin*	*Eyebrow*	*Pupil*	*White*	*Nostril*	*Lip*	*Background*	*Inner mouth*	*Average*
Pixel Acc	X	X	1.005	0.999	1.051	1.001	1.023	1.001	1.031	1.014	1.152	1.031
	X	○	1.014	1.004	1.125	1.021	1.059	0.962	1.021	1.008	1.132	1.038
	○	X	0.995	1.022	1.022	1.012	1.018	1.008	1.029	1.015	1.034	1.017
	○	○	1.003	1.017	1.127	1.020	1.050	0.960	1.064	1.023	1.022	1.032
IoU	X	X	0.992	1.009	0.930	0.987	1.015	1.047	1.040	1.016	1.627	1.074
	X	○	1.030	0.972	1.086	1.011	1.030	1.061	1.037	1.013	1.160	1.044
	○	X	1.011	1.004	1.052	0.979	1.242	1.134	0.982	0.961	1.271	1.071
	○	○	1.053	1.013	1.146	1.017	1.164	1.130	1.045	0.995	1.061	1.069

TABLE 5. Landmark extraction performance comparison of shallow network method.

MFB: Median frequency balancing, MAB: Median area balancing

Metrics	Methods		Classes									
	Class Weights	*Weighted Feature Map*	*Hair*	*Skin*	*Eyebrow*	*Pupil*	*White*	*Nostril*	*Lip*	*Background*	*Inner mouth*	*Average*
Pixel Acc	MFB	X	1.019	1.000	1.028	1.006	1.009	0.997	1.035	1.015	1.188	1.033
	MFB	○	1.031	0.994	1.016	1.029	1.043	0.983	1.080	1.000	1.255	1.048
	MAB	X	1.009	1.023	1.000	1.017	1.004	1.004	1.033	1.016	1.066	1.019
	MAB	○	1.020	1.008	1.018	1.028	1.034	0.982	1.126	1.015	1.133	1.040
IoU	MFB	X	1.031	1.016	1.021	1.001	1.087	1.486	1.023	1.003	2.682	1.261
	MFB	○	1.026	0.983	1.042	1.003	0.993	1.164	1.052	0.983	1.219	1.052
	MAB	X	1.050	1.011	1.155	0.993	1.329	1.608	0.966	0.949	2.095	1.240
	MAB	○	1.049	1.024	1.100	1.008	1.123	1.240	1.060	0.964	1.115	1.076

TABLE 6. Landmark extraction performance comparison of weighted feature maps.

MFB: Median frequency balancing, MAB: Median area balancing

Metrics	Methods		Classes									
	Class Weights	*Shallow Network*	*Hair*	*Skin*	*Eyebrow*	*Pupil*	*White*	*Nostril*	*Lip*	*Background*	*Inner mouth*	*Average*
Pixel Acc	MFB	X	0.958	1.028	0.872	0.958	0.881	0.868	0.899	1.007	0.882	0.928
	MFB	○	0.970	1.022	0.862	0.980	0.910	0.856	0.937	0.992	0.932	0.940
	MAB	X	0.967	1.033	0.933	0.977	0.911	0.834	0.890	1.001	0.868	0.935
	MAB	○	0.978	1.017	0.950	0.987	0.938	0.816	0.970	0.999	0.922	0.953
IoU	MFB	X	0.957	1.016	1.045	1.057	1.500	2.013	0.966	0.994	2.923	1.386
	MFB	○	0.953	0.982	1.067	1.059	1.371	1.577	0.993	0.974	1.328	1.145
	MAB	X	0.993	0.978	1.220	1.083	1.522	2.038	0.963	0.992	2.084	1.319
	MAB	○	0.992	0.991	1.162	1.100	1.285	1.572	1.056	1.007	1.109	1.142

(a) Original (b) Ground Truth (c) SegNet+MFB (d) SegNet+MAB

FIGURE 7. Landmark extraction results by MFB and MAB.

while IoU improved by more than 20% without the weighted feature maps, and more than 5% with the weighted feature maps. These improvements comes from the shallow network reducing misclassification. Figure 8 shows the effects of adding a shallow network. In the figure, original SegNet classified part of the hair as an eyebrow, while adding a shallow network eliminated this problem. This effect can be seen in other classes with very small pixel occupancy, such as the inner mouth. This kind of class shows both low pixel occupancy and low occurrence. So, if a model misclassifies such class, its IoU will drop dramatically. In contrast, if the class is classified correctly, its IoU will increase significantly. Figure 8 (b) shows the effect of adding a shallow network in extracting the inner mouth.

3) WEIGHTED FEATURE MAP

Classes that occupy a large area in the image are relatively easy to train, and in facial landmarks, the background, hair, and skin correspond to those classes. Weighted feature map assigns a larger weight to them to improve the classification

TABLE 7. Comparison of landmark extraction performance according to the Original SegNet.

MFB: Median frequency balancing, MAB: Median area balancing

Metrics	Methods			Classes									
	Class Weight	*Addition Network*	*Weighted Feature Map*	*Hair*	*Skin*	*Eyebrow*	*Pupil*	*White*	*Nostril*	*Lip*	*Background*	*Inner mouth*	*Average*
Pixel Acc	MFB	X	X	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	MAB	X	X	1.005	0.999	1.051	1.001	1.023	1.001	1.031	1.014	1.152	1.031
	MAB	○	X	1.013	1.022	1.051	1.018	1.027	1.005	1.065	1.030	1.228	1.051
	MAB	○	○	0.991	1.040	0.999	1.005	0.964	0.819	1.033	1.030	1.132	1.001
IoU	MFB	X	X	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	MAB	X	X	0.992	1.009	0.930	0.987	1.015	1.047	1.040	1.016	1.627	1.074
	MAB	○	X	1.042	1.020	1.074	0.980	1.350	1.685	1.005	0.964	3.408	1.392
	MAB	○	○	1.034	1.011	1.248	1.078	1.734	2.648	1.061	0.972	3.778	1.618

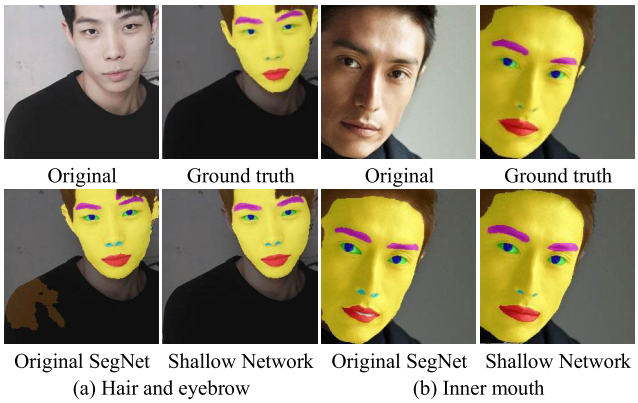


FIGURE 8. Landmark extraction quality comparison of original SegNet and SegNet with a shallow network.

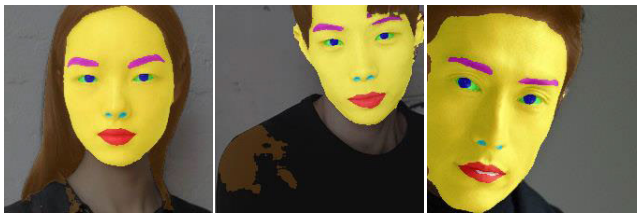


FIGURE 9. Facial landmark extraction by applying weighted feature maps to original SegNet.

performance of some ambiguous area, such as boundaries of classes. After all, it prevents the region of a small class from becoming too large by judging that the ambiguous part belongs to the large class, which greatly improved the IoU of small classes. Table 6 shows how pixel classification performance improved by using weighted feature maps. Overall, pixel accuracy dropped slightly, but IoU was greatly improved.

Figure 9 shows the result of applying the SegNet with our weighted feature map to the images in Figures 7 and 8. It can be seen that small landmarks such as eyebrows and nostrils were extracted more accurately than ever.

4) APPLYING MULTIPLE METHODS

So far, we showed that each of our three methods contributed to the precise extraction of facial landmarks. In any case, IoU was improved significantly compared to the original SegNet. In addition, Table 7 shows the performance improvement of

FIGURE 10. Facial landmark extraction using all three methods.

landmark extraction when class weight, extra shallow network and weighted feature map are applied to the original SegNet in that order. Our FLSNet, which used all three methods, achieved the best mIoU, while the pixel accuracy did not change much. Figure 10 shows the result of facial landmark extraction when using all three methods.

The problems that were observed in the original SegNet were almost solved, and the quality of landmark extraction was improved, especially for small landmarks. Figure 11 shows 20 different sample images and their ground truth images used in the experiments, and Figure 12 presents landmark extraction results by applying the three methods sequentially.

In the figure, we can see that the landmark extraction quality was improved gradually as we applied those methods sequentially.

B. COMPARISION WITH OTHER METHODS

In this section, we compare our model with six well-known semantic segmentation methods in terms of IoU. The methods we considered in this experiment are FCNs, SegNet, PSPNet, Deeplab v3+ with Xception, Deeplab v3+ with InceptionResNetv2, and HRNet. For evaluation, we trained other semantic segmentation networks using the semantic segmentation dataset that we constructed in the previous experiments. Comparative evaluation were done under the same conditions except the mini-batch size because of the memory limitation. Table 8 summarizes the results. As shown in the table, all methods showed very good performance for the large-sized landmarks such as hair and skin. However, for small-sized landmarks, our method showed overwhelming performance, and achieved the highest mIoU.

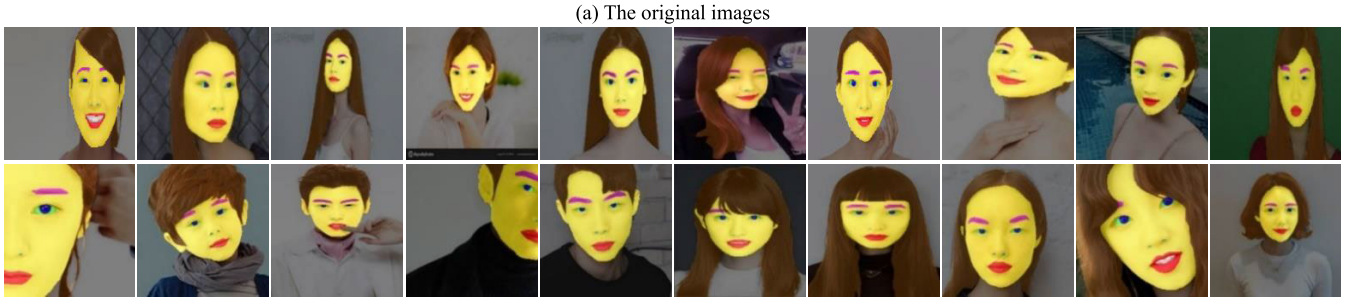
TABLE 8. Comparison of seven methods in terms of IoU.

Models		Classes									
Methods	*Backbones*	*Hair*	*Skin*	*Eyebrow*	*Pupil*	*White*	*Nostril*	*Lip*	*Background*	*Inner mouth*	*Average*
FCNs [23]	Vgg 16 [19]	0.845	0.885	0.659	0.522	0.335	0.212	0.812	0.914	0.301	0.610
SegNet [17]	Vgg 16 [19]	0.883	0.913	0.664	0.822	0.456	0.263	0.822	0.943	0.209	0.664
PSPNet [26]	ResNet-101 [42]	0.901	0.913	0.641	0.802	0.459	0.296	0.781	0.933	0.260	0.665
Deeplab v3+ [29]	Xception [31]	0.902	0.901	0.589	0.861	0.328	0.149	0.788	0.941	0.422	0.653
Deeplab v3+ [29]	InceptionResNetv2 [43]	0.890	0.900	0.671	0.818	0.447	0.266	0.798	**0.949**	0.477	0.691
HRNet [33]	HRNet-W48+ [33]	**0.932**	**0.930**	0.704	0.834	0.531	0.309	0.829	0.937	0.487	0.718
Our FLSNet	Vgg16 [19]	0.913	0.922	**0.828**	**0.886**	**0.790**	**0.697**	**0.873**	0.916	**0.790**	**0.846**

TABLE 9. Comparison with other model using the dataset in [22].

Methods	Classes							
	Skin	*Eyebrows*	*eyes*	*Nose*	*Upper Lip*	*Inner Mouth*	*Lower Lip*	*Average*
Jackson et al. [22] with Guided by Detected	0.935	0.862	0.759	0.534	0.512	0.569	0.393	0.652
Jackson et al. [22] with Guided by Ground truth	**0.943**	**0.872**	**0.819**	0.601	0.556	0.596	0.420	0.687
Our FLSNet	0.918	0.720	0.695	**0.854**	**0.761**	**0.861**	**0.803**	**0.802**

(a) The original images



(b) Corresponding ground truth images

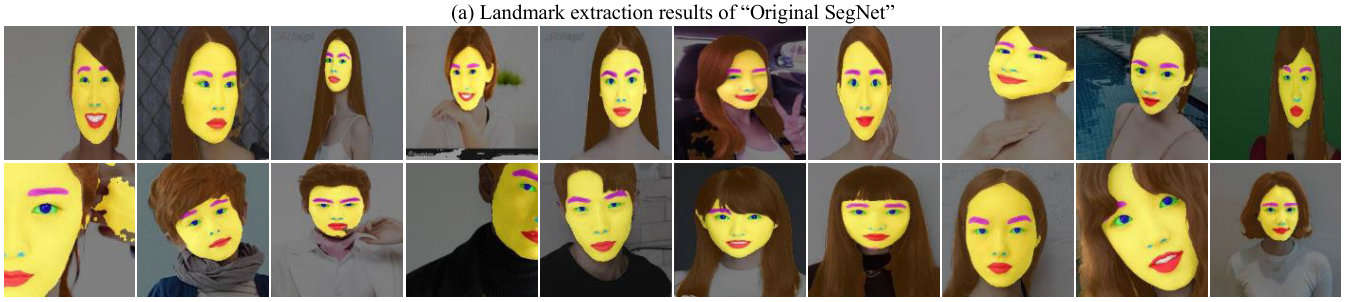
FIGURE 11. 20 samples used in experiments and their ground truth images.

C. EVALUATION WITH OTHER DATASET

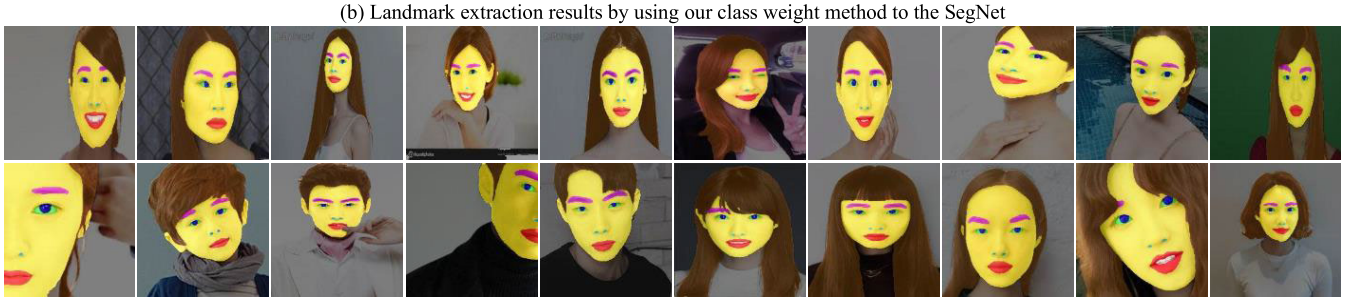
As mentioned before, there is no public facial landmark dataset. But, Jackson and colleagues constructed a dataset for semantic segmentation based on a public facial dataset that includes 68 key points for each facial image. For each facial landmark, they created a closed shape by connecting the points that represents the landmark.

Using the dataset, they proposed FCNs based facial landmark extraction method. The main idea of the method is offering facial landmark points to the semantic segmentation model as a guideline. Unfortunately, as many face images in our datasets do not include landmark points, they are not enough for training. Instead, we trained our model using the dataset used in [22]. Table 9 shows the comparison results. The table indicates that for eyebrows and eyes, Jackson's method showed better performance, and for skin, both Jackson's method and our method showed equally good performance. For the other classes, our method showed better performance. Overall, our method outperforms Jackson's method in terms of mIoU for most facial landmarks.

(a) Landmark extraction results of "Original SegNet"



(b) Landmark extraction results by using our class weight method to the SegNet



(c) Landmark extraction results by adding our shallow network to (b)

(d) Landmark extraction results by using all three methods

FIGURE 12. Facial landmark extraction result for 20 sample images. (a): Original SegNet (b): Training using our class weights (c): Adding a shallow network to (b) (d): Using weighted feature maps to (c).

V. CONCLUSIONS

The aim of the work described in this paper was to develop algorithms for robust facial landmark extraction. To achieve this aim, we first constructed a facial landmark dataset for semantic segmentation. Then, we developed a semantic segmentation architecture for robust facial landmark extraction. Next, we introduced a balancing scheme to coordinate pixel imbalance in the training process, and a data augmentation method to make the model robust in the multi-scale face. We also improved the quality of pixel classification by applying weights to the feature maps from the Softmax layer. In the experiments, using several semantic segmentation models, we demonstrated that our approach is effective for the quantitative detection of facial landmarks. The metric which we investigated, mean intersection over union (mIoU) of the facial landmarks, especially the IoU of small landmarks such as eye pupils and whites, improved significantly. Qualitatively, it is apparent that the landmarks extracted by our approach are cleaner and more accurate than the originals. Our approach has two main contributions. The first one is an improved semantic segmentation method: adding a shallow network and using our novel class weights in training, introducing a classification method by applying weights to feature maps, and producing a facial landmark dataset for semantic segmentation. The second contribution is improving human face-based technology such as individual identification, by merging our approach and previous research into extracting facial landmarks.

In future research, we will investigate the validity of our approach using facial landmark datasets as well as public data. In this work, we used the weighted feature maps separately from the training process. However, it is possible that performance could be improved by using the maps directly to train the semantic segmentation model. We will, therefore, research the use of weighted feature maps for training.

REFERENCES

[1] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[3] H. Kim, H. Kim, and E. Hwang, "Real-time shape tracking of facial landmarks," *Multimedia Tools Appl.*, pp. 1–19, Nov. 2018, doi: 10.1007/s11042-018-6814-7.

[4] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image Vis. Comput.*, vol. 47, pp. 27–35, Mar. 2016.

[5] Q. Hou, J. Wang, L. Cheng, and Y. Gong, "Facial landmark detection via cascade multi-channel convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 1800–1804.

[6] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 160–169.

[7] 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.

[8] 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.

[9] H. Kim, H. Kim, and E. Hwang, "Real-time facial feature extraction scheme using cascaded networks," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2019, pp. 1–7.

[10] A. Juhong and C. Pintavirooj, "Face recognition based on facial landmark detection," in *Proc. 10th Biomed. Eng. Int. Conf. (BMEiCON)*, Aug. 2017, pp. 1–4.

[11] W. Abdalmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, R. Nevatia, and G. Medioni, "Face recognition using deep multi-pose representations," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[12] I. Tautkute, T. Trzcinski, and A. Bielski, "I know how you feel: Emotion recognition with facial landmarks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1878–1880.

[13] F. Khan, "Facial expression recognition using facial landmark detection and feature extraction via neural networks," 2018, *arXiv:1812.04510*. [Online]. Available: http://arxiv.org/abs/1812.04510

[14] H. Kim, J. Park, H. Kim, and E. Hwang, "Facial landmark extraction scheme based on semantic segmentation," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Jan. 2018, pp. 1–6.

[15] H. Kim, J. Park, H. Kim, E. Hwang, and S. Rho, "Robust facial landmark extraction scheme using multiple convolutional neural networks," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3221–3238, Feb. 2019.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[17] 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.

[18] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," 2015, *arXiv:1505.07293*. [Online]. Available: http://arxiv.org/abs/1505.07293

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[20] X. Tang, F. Guo, J. Shen, and T. Du, "Facial landmark detection by semi-supervised deep learning," *Neurocomputing*, vol. 297, pp. 22–32, Jul. 2018.

[21] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3486–3496.

[22] A. S. Jackson, M. Valstar, and G. Tzimiropoulos, "A CNN cascade for landmark guided semantic part segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 143–155.

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

[24] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015.

[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[27] 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*. [Online]. Available: http://arxiv.org/abs/1412.7062

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "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, Apr. 2018.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[30] 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. 801–818.

[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[32] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Trans. Image Process.*, early access, Jan. 25, 2019, doi: 10.1109/TIP.2019.2895460.

[33] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*. [Online]. Available: http://arxiv.org/abs/1904.04514

[34] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.

[35] M. Svanera, U. R. Muhammad, R. Leonardi, and S. Benini, "Figaro, hair detection and segmentation in the wild," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 933–937.

[36] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 679–692.

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

[38] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, Dec. 2013.

[39] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting CRFs with Boltzmann machine shape priors for image labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2019–2026.

[40] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-Wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 397–403.

[41] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[42] 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.

[43] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.