Artificial Intelligence In Medicine 145 (2023) 102679

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Facial wrinkle segmentation using weighted deep supervision and semi-automatic labeling☆

Semin Kim *, Huisu Yoon, Jongha Lee, Sangwook Yoo

AI R&D Center, Lululab Inc., 318, Dosan-daero, Gangnam-gu, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Wrinkle segmentation
Wrinkle detection
Deep supervision
Semi-automatic labeling
Deep learning
Retinal vessel segmentation
U-Net

ABSTRACT

Facial wrinkles are important indicators of human aging. Recently, a method using deep learning and a semi-automatic labeling was proposed to segment facial wrinkles, which showed much better performance than conventional image-processing-based methods. However, the difficulty of wrinkle segmentation remains challenging due to the thinness of wrinkles and their small proportion in the entire image. Therefore, performance improvement in wrinkle segmentation is still necessary. To address this issue, we propose a novel loss function that takes into account the thickness of wrinkles based on the semi-automatic labeling approach. First, considering the different spatial dimensions of the decoder in the U-Net architecture, we generated weighted wrinkle maps from ground truth. These weighted wrinkle maps were used to calculate the training losses more accurately than the existing deep supervision approach. This new loss computation approach is defined as weighted deep supervision in our study. The proposed method was evaluated using an image dataset obtained from a professional skin analysis device and labeled using semi-automatic labeling. In our experiment, the proposed weighted deep supervision showed higher Jaccard Similarity Index (JSI) performance for wrinkle segmentation compared to conventional deep supervision and traditional image processing methods. Additionally, we conducted experiments on the labeling using a semi-automatic labeling approach, which had not been explored in previous research, and compared it with human labeling. The semi-automatic labeling technology showed more consistent wrinkle labels than human-made labels. Furthermore, to assess the scalability of the proposed method to other domains, we applied it to retinal vessel segmentation. The results demonstrated superior performance of the proposed method compared to existing retinal vessel segmentation approaches. In conclusion, the proposed method offers high performance and can be easily applied to various biomedical domains and U-Net-based architectures. Therefore, the proposed approach will be beneficial for various biomedical imaging approaches. To facilitate this, we have made the source code of the proposed method publicly available at: https://github.com/resemin/WeightedDeepSupervision.

1. Introduction

Facial wrinkles are major indicators for estimating human age [1] and identifying human emotions [2]. Numerous researchers have proposed facial wrinkle segmentation methods, and cosmetics companies continue to launch various types of wrinkle treatments. Facial wrinkle segmentation is an important area of research for preventing facial aging.

Most wrinkle segmentation approaches have been developed based on Hessian or Gabor filters [3–6]. The components of the Hessian matrix are partial second-order derivatives. The eigenvalues of a 2D image can be used to characterize the key features of each pixel. The Gabor filter is a linear filter which a Gaussian kernel modulated by a sinusoidal function. Through convolution with a Gabor filter, the magnitude and direction of a component of a particular frequency can be emphasized in the corresponding image. Applying pre/post-processing filters to a Gabor-filtered image enables the detection of facial wrinkles [7,8]. Despite having reliable performance, the aforementioned approaches were considered to have limitations because they depend on geometric assumptions about wrinkles that can limit performance in particular images. In addition, the filter parameters must be fine-tuned [7] to improve the performance of wrinkle segmentation, because the optimal

☆ This article belongs to Special issue: CBMS22 Mining healthcare

* Corresponding author.
E-mail addresses: sm.kim@lulu-lab.com (S. Kim), hs.yoon@lulu-lab.com (H. Yoon), jongha.lee@lulu-lab.com (J. Lee), sangwook.yoo@lulu-lab.com (S. Yoo).

https://doi.org/10.1016/j.artmed.2023.102679
Received 15 January 2023; Received in revised form 28 July 2023; Accepted 3 October 2023
Available online 6 October 2023

parameters for each face image can differ.

In the field of image recognition research, studies have primarily focused on machine learning and feature fusion [9–12]. Recently, large-scale training datasets have been extensively used with deep-learning techniques. Extensive studies based on deep learning have been conducted in the area of biomedical imaging. Many research studies have been proposed to utilize deep learning for segmentation, especially in medical imaging modalities such as MRI [13], CT [14], and ultrasound [15]. Notably, segmentation research focusing on human organs has been introduced, targeting the heart [16], brain tissues [17,18], and liver [19]. In contrast, there have also been research endeavors aimed at segmenting bones from various organs [20–22]. Additionally, methods for segmenting small ocular vessels and retinas have been proposed [23–26]. Furthermore, deep learning-based segmentation research has been conducted on histopathological or pathological images observed through microscopes [27,28]. Indeed, various approaches using deep learning have been suggested to perform segmentation of skin diseases on the human face and skin [29–32].

Basically, these datasets are typically labeled manually for specific objects. However, because the shape and thickness of facial wrinkles vary, it is challenging to precisely label wrinkles manually. Therefore, because of the difficulty in generating wrinkle labels, there has been no adequate research on wrinkle segmentation for the entire face based on deep learning. A semi-automatic labeling strategy utilizing deep learning was proposed in a recent study [33] to mitigate the expenses associated with labeling facial wrinkles. Rough labels of facial wrinkles and an adaptive thresholding technique [34] were used to create wrinkle labels (ground truth). Using these wrinkle labels, a deep learning-based wrinkle segmentation model was trained, which outperformed more traditional image-processing-based techniques in terms of performance.

Despite the application of semi-automatic labeling, the detection of wrinkles on faces requires further improvement due to the lower segmentation performance compared to other biomedical domains. To address this issue, this study employed deep supervision [35–37] to improve wrinkle segmentation performance. Recently, numerous methods in the field of biomedical image recognition have incorporated deep supervision to enhance their performance. During the training phase, deep supervision extracted feature maps from each decoder of the training model and calculated multiple losses. In general, the decoders in the training model have smaller spatial dimensions owing to downsampling. Thus, to make them the same size as the ground truth, upsampling layers were utilized. However, upsampling cannot accurately depict thin objects, such as facial wrinkles. In particular, as the spatial dimension decreases, it becomes more difficult to depict the detailed shapes of facial wrinkles, and the training loss becomes more inaccurate.

To address this issue, we propose a new technique based on the deep supervision of wrinkle segmentation. We found that when calculating the loss for facial wrinkle inference, it is necessary to compute the loss differently for downsampled decoders. Thus, the proposed method employs weighted wrinkle maps (*WWM*) to calculate training losses from downsampled decoders more precisely. The *WWM* was generated primarily from the ground truth by average pooling and upsampling. The use of *WWM* as weight factors during the computation of training losses was implemented to reduce incorrect losses. Loss computation utilizing *WWM* is defined as weighted deep supervision.

The proposed method was compared to conventional image-processing methods and deep supervision. The dataset was acquired using a specialized skin analysis device [38] and wrinkle labeling was accomplished using a semi-automatic technique [33]. The proposed method demonstrated significantly superior performance compared with conventional image-processing techniques. Furthermore, it produced better performance than deep supervision. In addition, we measured the consistency of the labels generated by humans and those generated by semi-automatic labeling for the same images. In our experiment, three labelers generated wrinkle labels for the same images, and we applied a semi-automatic labeling technique to generate wrinkle labels. We calculated and compared the consistency of each generated wrinkle label based on this correlation. The correlation coefficient of the semi-automatic labeling technique was much higher, indicating that using semi-automatic labeling was much more effective in generating consistent wrinkle labels than labeling directly by humans. Finally, we analyzed the scalability and limitations of the proposed method.

The main novelty of this paper can be summarized as follows:

- *Weighted wrinkle map*: We propose a weighted wrinkle map that considers the spatial dimensions of the decoder layers when computing losses.
- *Weighted deep supervision*: We propose a new deep-learning framework that uses *WWM* to compute training losses.
- *Consistency of wrinkle labels*: We show a better performance of semi-automatic labeling than that of a human labeling job for wrinkle segmentation.

The rest of this paper is organized as follows: In Section 2, we provide a brief introduction to wrinkle segmentation models based on traditional image processing and deep supervision. In Section 3, the proposed weighted deep supervision method is described in detail. In Section 4, we compare the wrinkle detection performance of our proposed method with that of existing methods and conduct experiments to compare the consistency of human labeling and semi-automatic labeling. Section 5 discusses the scalability and limitations of the proposed method. Finally, we present our conclusions in Section 6.

2. Related works

2.1. Wrinkle segmentation based on image processing

The eigenvalues of the Hessian matrix represent the magnitude of displacement along the respective eigenvectors. Using this knowledge, the Frangi filter [3] was developed to detect the vesselness of magnetic resonance angiography. The Frangi filter was incorporated into a Hybrid Hessian Filter (*HHF*) to approximate the wrinkle-related structure on the horizontal gradient map [4]. Following a series of additional processes, such as thresholding, the wrinkles were refined. Hessian Line Tracking included the *HHF* as one of the phases to exploit wrinkle connectivity [5] as a follow-up study. However, the method of extracting wrinkles using gradients in only one direction is ineffective for extracting wrinkles in multiple dimensions.

The Gabor filter response highlights the signal components of the corresponding orientations and frequencies. In Cula et al.'s method [6], Gabor filter was incorporated to extract wrinkles using a local orientation map generated based on the gradient information of the target image. Batool et al. extracted wrinkle candidate structures by passing them through a Gabor filter bank and selecting the maximum value for each pixel [8]. To obtain good results, appropriate Gabor filter parameters should be set for individual images. It was difficult to determine the optimal filter parameters for input images, which varied significantly depending on individual differences, shooting conditions, and image resolution.

In addition, these image-processing methods do not distinguish between hair, eyebrows, and eyes. Therefore, additional pre-processing is required.

2.2. Deep supervision

Deep supervision calculates multiple losses in multiple layers of a training model. These losses are used to update the training model, which exhibits better performance than a single loss. Deep supervision is typically applied in U-Net [39] structures such as M-Net [35], AG-Net [36], U-Net3+ [37]. M-Net and AG-Net computed four losses from the

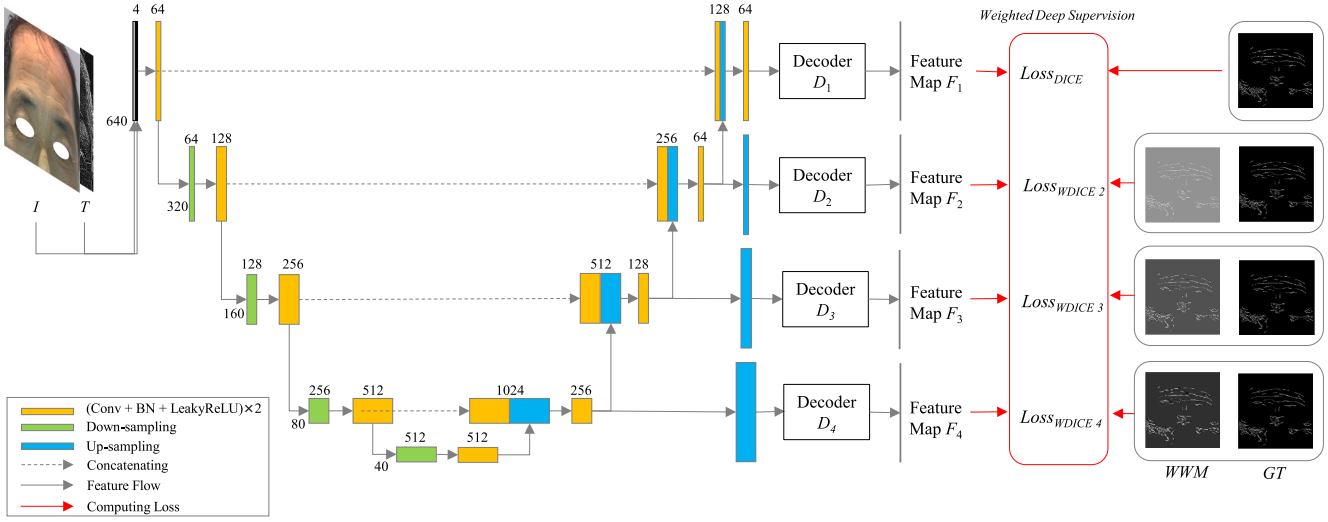

Fig. 1. Overview of the proposed wrinkle segmentation model. An original image I and a texture map T are concatenated and used as input. 'Conv' is the convolutional layer, and 'BN' is the batch normalization layer, and Leaky ReLU is the activation function. The numbers at the bottom left of the features indicate the width or height size, and the numbers above the features indicate the number of channels. On the right, there are ground truth and weighted wrinkle maps. The red boxes represent the weighted deep supervision proposed in this paper. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

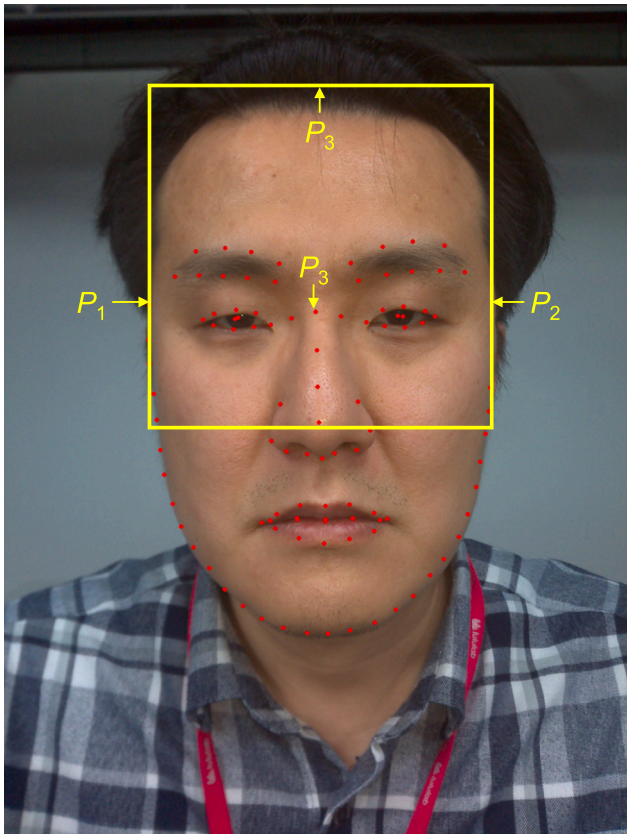

Fig. 2. Example of cropping a face using landmark points (red). First, three points (P_1, P_2, and P_3) are selected, and P_4 is estimated from the three points. Then, two points (x_1, y_1) and (x_2, y_2) are computed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

four decoders, with the exception of the bottom decoder for retinal vessel segmentation. U-Net3+ computed the losses for all decoders and updated a training model for liver and spleen segmentation in the CT images. These approaches demonstrated better performance than a single loss in their experiment. However, deep supervision has primarily been applied to downsampled decoders. As a result, the loss may be inaccurate for thin objects, such as wrinkles.

3. The proposed method

This section provides a detailed description of the proposed facial wrinkle segmentation technique. First, we present an outline of the overall structure of the proposed method. The procedure for pre-processing facial images to detect facial creases is described below. Subsequently, a method for creating the proposed weighted wrinkle map is described. The final section describes the procedure for calculating the losses using the proposed weighted deep supervision.

3.1. Overview of the wrinkle segmentation model

Fig. 1 shows the overall structure of the proposed wrinkle segmentation based on U-Net [33,40]. Furthermore, it is based on deep supervision, which extracts feature maps from each decoder and computes multiple losses. However, except for feature map F_1, the losses were calculated using the proposed weighted deep supervision.

3.2. Pre-processing

Facial wrinkle segmentation has mainly been studied for the forehead and areas around the eyes [41–43]. The forehead contains many expressive muscles that move frequently, and the areas around the eyes have thin skin and expressive muscles that result in wrinkles. By focusing on these two areas, we were able to estimate the overall aging of the face. This study also presents the development of facial wrinkle segmentation techniques for these specific areas. To detect wrinkles around the forehead and eyes, we applied Jin et al.'s [44] method to detect landmark points. Landmarks are a method of extracting key-points, such as eyes, eyebrows, nose, mouth, and jawline. In Fig. 2, the red dots represent the landmarks, each with coordinates (x, y). The leftmost point was defined as P_1, the rightmost point as P_2, and the middle point between the eyes as P_3. However, because there was no landmark on the forehead, a virtual landmark was created and defined as P_4. The y-coordinate of P_4 is the same as y_1 in Table 1, and is

Table 1
Methods for calculating coordinates to crop the face.

Coordinate	Definition
x_1	x of P_1
x_2	y of P_2
y_1	x of $P_3 - (x_2 - x_1) \times 0.66$
y_2	$y_1 + (x_2 - x_1)$

calculated using P_1, P_2, and P_3. Finally, the two points (x_1, y_1) and (x_2, y_2), representing the green box drawn in Fig. 2 are calculated as shown in Table 1.

3.3. Semi-automatic wrinkle labeling

Because facial wrinkles are quite diverse in shape and length and their boundaries are ambiguous, it is difficult for a labeling annotator to label them by hand. Therefore, as shown in Fig. 3, a semi-automatic labeling technique [33] was applied to generate the wrinkle labels. As shown in Fig. 3, a labeling annotator creates a rough wrinkle annotation map. This map is converted into a binary mask M. Then, a texture map T is extracted from the original image I using Eq. (1).

$$T(x,y) = \left(1 - \frac{I(x,y)}{1 + I_{G(\sigma)}(x,y)}\right) \times 255, \tag{1}$$

where G is a Gaussian kernel, σ is a sigma value, and $I_{G(\sigma)}$ is a Gaussian filtered image, x and y are coordinates. Then, to remove the non-wrinkled texture from the texture map, we used Eq. (2), as follows:

$$T'(x,y) = \begin{cases} T(x,y), & \text{if } M(x,y) > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

Finally, an adaptive thresholding method [34] was applied to generate the ground truths.

3.4. Weighted wrinkle map

This subsection explains the generation of the WWM using Algorithm 1. Decoders D_1, D_2, and D_3 in Fig. 1 have reduced spatial dimensions owing to the downsampling. Therefore, it was necessary to adjust the weights of each area in the GT by considering the reduced spatial dimensions. Wrinkle detection performance can be improved by calculating the loss using the adjusted weights. The images in the first column of Fig. 4 show the results of Steps 1 and 2 of Algorithm 1. The scale factor s represents a reduction in the spatial dimension at a ratio of $1/s$. Thus, the representation of the fine details of the feature map was reproduced using the ground truth through a process of downsampling and upsampling, resulting in Steps 3 and 4. Therefore, the role of *WWM* is to decrease the weights of areas where wrinkles cannot be properly represented through upsampling and to maintain high weights for areas where wrinkles are indispensable.

Algorithm 1. Generating a weighted wrinkle map (*WWM*).

Input: A ground truth image $GT \in \mathbb{R}^{h \times w}$, scale factor s
Output: A weighted wrinkle map $WWM \in \mathbb{R}^{h \times w}$
Step 1. Compute the means of sub-blocks in GT
  a. GT is partitioned into a $h_d \times w_d$ grid format, where $h_d = h/s$, $w_d = w/s$
  b. For each grid, the average values are calculated to generate $WWM_{down} \in \mathbb{R}^{h_d \times w_d}$.
Step 2. Up-sample WWM_{down} using nearest interpolation to generate $WWM \in \mathbb{R}^{h \times w}$.
Step 3. Calculate the mean value of WWM and replace zero values with the computed mean value.
Step 4. Set the values of WWM to 1 at the same positions as where the values of GT are 1.

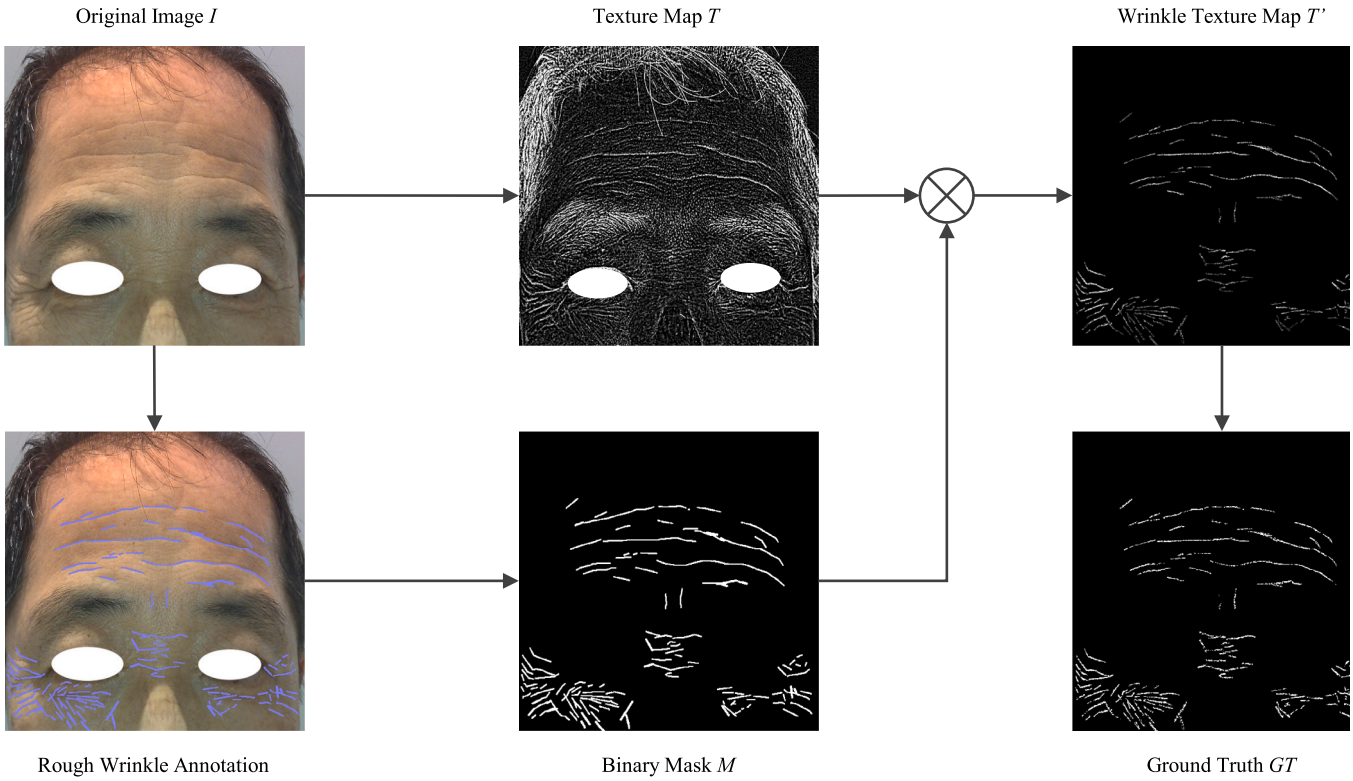

Fig. 3. Overview of semi-automatic labeling [33]. First, the binary mask M is generated by roughly labeling wrinkle areas. The proposed texture map T is created from the original image I. Then, non-wrinkle textures are removed by multiplying with T and M, and the wrinkle texture map T' is created. Finally, the ground truth GT is generated by adaptive thresholding from T'.

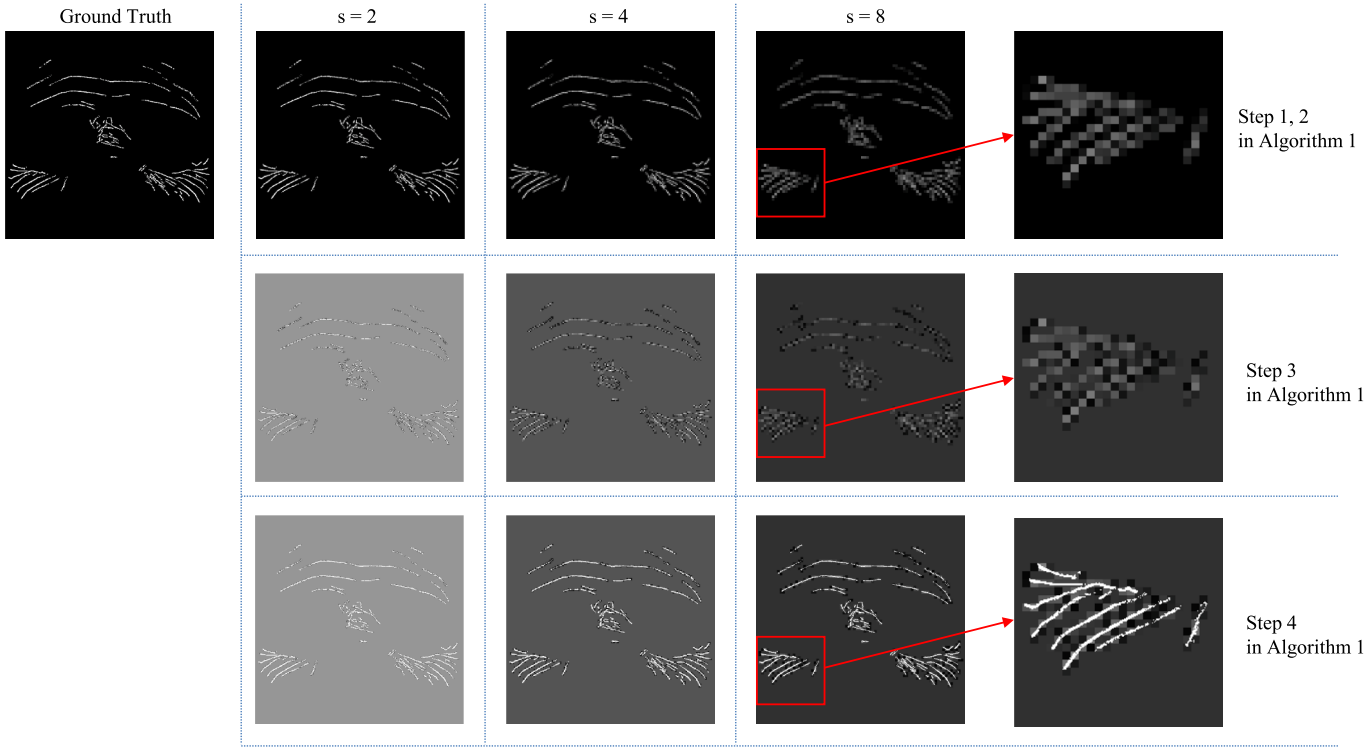

Fig. 4. Example of the weighted wrinkle maps in Algorithm 1.

3.5. Loss functions

Two loss functions were used to train the proposed wrinkle segmentation model. We used the DICE coefficient loss for the last decode layer. The DICE coefficient loss is defined as

$$Loss_{DICE\ n} = 1 - 2\frac{\sum_i p_{n,i} q_i}{\sum_i p_{n,i} + \sum_i q_i}, \tag{3}$$

where $p_{n,i}$ is ith value of nth predicted output, such as F_n and q is the ground truth. For the remaining decoder layers, we propose a weighted DICE coefficient loss defined as

$$Loss_{WDICE\ n} = 1 - 2\frac{2 \times \sum_i p_{n,i} q_i w_{n,i}}{\sum_i p_{n,i} w_{n,i} + \sum_i q_i w_{n,i}}, \tag{4}$$

where $w_{n,i}$ is ith value of nth *WWM*. Thus, the final loss can be computed as

$$Loss = Loss_{DICE\ 1} + \sum_{i=2}^{4} Loss_{WDICE\ i}, \tag{5}$$

where i is ith feature map, or *WWM*. In this study, the method of training the wrinkle segmentation model based on the above loss function was defined as weighted deep supervision.

3.6. Learning rate scheduler

Learning rate schedulers are widely used to train deep learning models with high performance. In the proposed method, a cosine annealing learning rate scheduler (CALRS) [45] was used in the training phase. CALRS is defined as below:

$$\gamma_t = \gamma^i_{min} + \frac{1}{2}(\gamma^i_{max} - \gamma^i_{min})\left(1 + cos\left(\frac{T_{cur}}{T_i}\pi\right)\right), \tag{6}$$

where $\gamma^i_{min} + \gamma^i_{max}$ are the range for the learning rate in the ith run, and T_{cur} is the number of epochs that have progressed from the last restart, and t is the batch iteration. We used CALRS to adjust learning rate to update the weights of the proposed wrinkle segmentation model.

4. Experiments

4.1. Experimental Environments

To evaluate the performance of the proposed weighted deep supervision method, 300 facial images were acquired using a specialized skin diagnosis device, Lumini KIOSK v2 [38]. Facial images were obtained from various locations with different lighting and color temperatures, making this database quite challenging. All acquired images included a person's face facing forward, as shown in Fig. 2. We performed 6-fold cross-validation experiments on 300 images to obtain the experimental results. The acquired image was 1280 (H) × 960 (W), and all images included a frontal face. All the acquired images were cropped and resized to 640 (H) × 640 (W) by pre-processing using landmark detection [44], as described in Section 3.2. Subsequently, ground truths (GT) for wrinkles were obtained using a semi-automatic labeling technique and expert consultation. To create texture map T, the sigma value was set to 5, and the size of the Gaussian kernel was set to 21 × 21.

In the training phase, data augmentation techniques for scaling, shifting, rotating, brightness, color changing, and flipping were used randomly when the training images were loaded. Adam [46] was applied as the optimizer to update the weights of the proposed wrinkle segmentation model. The period of the cosine-annealing learning rate scheduler (CALRS) [45] and the maximum epoch of the training phase were set to 200. The initial learning rate was set to 0.01, and the minimum learning rate was set to 0.000001. The weight decay was defined based on L_2-norm with a value of 0.0001, and the batch size was 4. All

Table 2
Statistical comparison of facial wrinkle segmentation.

Metric	JSI (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
U-Net	42.30 ± 1.05	98.57 ± 0.13	62.92 ± 0.82	99.24 ± 0.10
U-Net + DS	42.98 ± 0.81	98.60 ± 0.12	63.47 ± 0.76	99.19 ± 0.08
U-Net + WDS	44.35 ± 0.94	98.67 ± 0.12	63.84 ± 0.40	99.25 ± 0.08

experimental results were obtained using PyTorch and an NVIDIA RTX 3090 GPU on Ubuntu 20.04.

4.2. Metrics

Several metrics were chosen to evaluate the proposed method using previous approaches. The Jaccard Similarity Index (JSI), accuracy, specificity, and sensitivity were evaluated. The JSI has been frequently employed in previous wrinkle segmentation methods [4–6,33] and is defined as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{7}$$

where A is a prediction and B is the ground truth. Accuracy, sensitivity, and specificity are defined as follows [47]:

$$Accuracy = \left(\frac{TN + TP}{TN + TP + FP + FN}\right), \tag{8}$$

$$Sensitivity = \left(\frac{TN}{TN + FN}\right), \tag{9}$$

$$Specificity = \left(\frac{TN}{TN + FN}\right). \tag{10}$$

where TP represents true positives, TN is true negatives, FP is false positives, and FN represents false negatives.

4.3. Performance of wrinkle segmentation

We selected U-Net [33,41] as the baseline and compared the wrinkle segmentation performance by applying weighted deep supervision and deep supervision separately with 6-fold cross-validation. Table 2 presents a statistical summary of each method based on the averages and variances of the metrics. The weighted deep supervision demonstrated the best performance among the four metrics. In the case of JSI, although the use of deep supervision resulted in a 0.68 % improvement, the use of weighted deep supervision led to a 2.05 % improvement. This is because the discriminative power for wrinkles increased, leading to an improvement in accuracy, as well as sensitivity and specificity. This demonstrates that weighted deep supervision is more effective than conventional deep supervision in enhancing wrinkle segmentation performance. Fig. 5 shows examples of the results obtained using each method.

In addition, we compared our proposed method with traditional wrinkle detection methods that use image-processing techniques, such as Hessian [4] or Gabor [8]. However, when attempting to detect wrinkles on the entire face using traditional methods, other features, such as eyes, eyebrows, and hair, are also detected as wrinkles, resulting in very low performance. Therefore, to conduct a fair experiment, we measured the performance only on the forehead and around the eyes in the facial images. First, we selected one from the 6-fold split training set and manually cropped the forehead and eye areas. We then filled the non-skin areas with black to prevent traditional methods from detecting eyebrows or hair as wrinkles. Table 3 presents the wrinkle detection results for each method on the forehead and around the eyes. Our proposed method showed significantly higher performance than traditional methods. Figs. 6 and 7 show examples of the results of each method.

4.4. Inefficiency of deep supervision for wrinkle segmentation

Deep supervision has been extensively employed in biomedical image recognition. However, deep supervision is inefficient for thin objects such as wrinkles. The two feature maps F_1 and F_4 are shown in Fig. 8. These features were extracted from decoders D_1 and D_4 as

Table 3
Comparison of JSI performance. DS is deep supervision, WDS is weighted deep supervision.

Part	Hessian	Gabor	U-Net	U-Net + DS	U-Net + WDS
Forehead	0.173	0.142	0.472	0.469	0.486
Eyes	0.181	0.159	0.443	0.455	0.460
Average	0.177	0.151	0.458	0.462	0.473

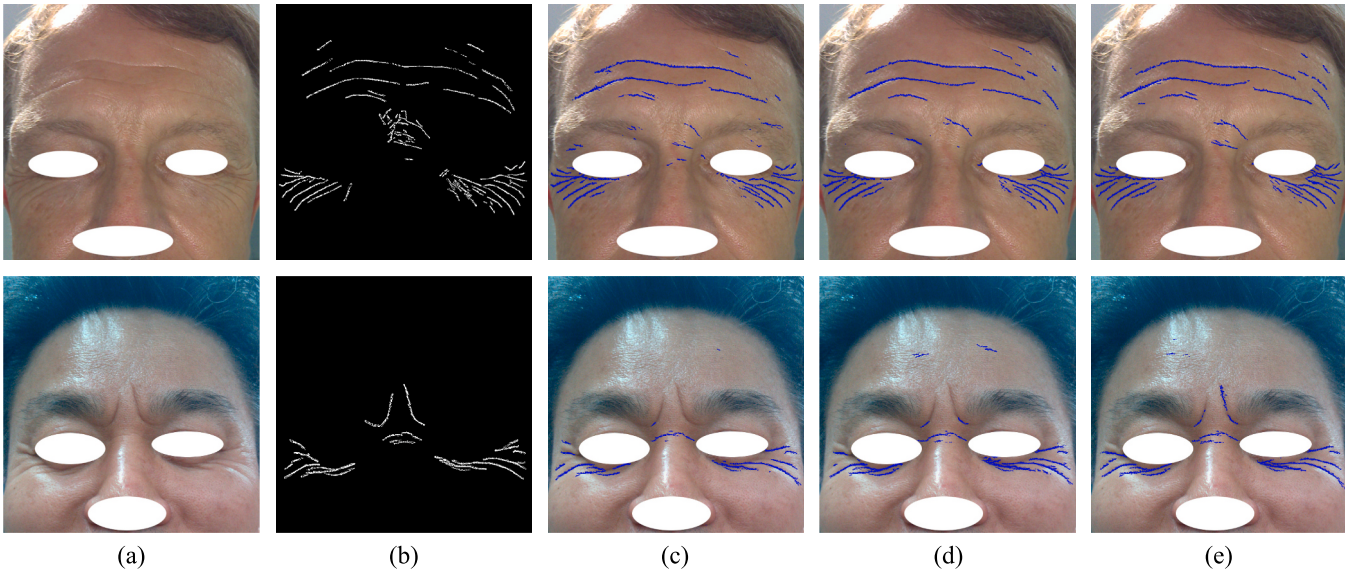

Fig. 5. Comparison of wrinkle results on entire face. (a) is an original image, (b) is a ground truth, (c) is a result of U-Net without deep supervision, (d) is a result of U-Net with deep supervision, and (e) is a result of the proposed method, weighted deep supervision.

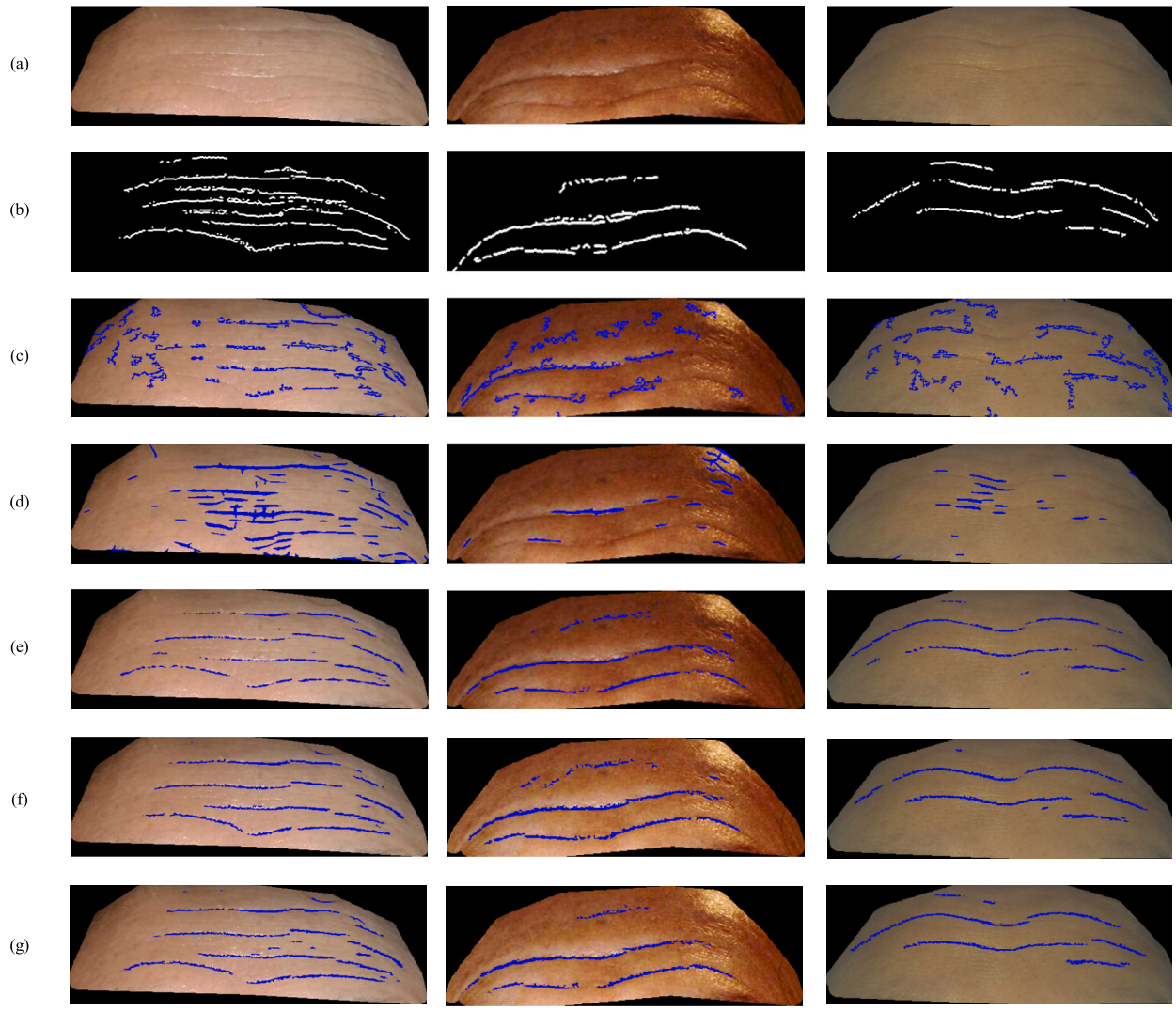

Fig. 6. Comparison of wrinkle results on foreheads. (a) is an original image, (b) is a ground truth, (c) is a result of Hessian approach [4], (d) is a result of Gabor filter approach [8], (e) is a result of U-Net without deep supervision, (f) is a result of U-Net with deep supervision, and (g) is a result of the proposed method, weighted deep supervision.

illustrated in Fig. 1. In the case of F_4, the spatial dimension decreased to 1/8, resulting in a blurred image when upsampled again by a factor of eight. Therefore, it is difficult to precisely represent the shape of the wrinkles. In contrast, F_1 was more specific than F_4. However, this may not be appropriate because the training loss for F_4 is consistently higher than that for F_1, which may lead to a larger proportion of F_4 in the training phase. As the final performance is determined through F_1, this is not correct.

To examine the results further from a loss perspective, we compared the training losses for deep and weighted deep supervision. In Fig. 9, the training loss of $Loss_{DICE4}$ under deep supervision is relatively high compared to that of $Loss_{DICE1}$. Thus, the loss of $Loss_{DICE4}$ can significantly affect the overall model training. However, because F_1 is the final prediction result, its loss should have more weight in model training. In the case of weighted deep supervision, $Loss_{DICE1}$ was maintained at a higher level than $Loss_{WDICE4}$. Therefore, $Loss_{DICE1}$ had a greater influence on training. As a result, it can be observed that $Loss_{DICE1}$ of weighted deep supervision is lower than that of deep supervision. Hence, it can be concluded that the training performance of weighted deep supervision is superior.

Furthermore, we explain the main reason why $Loss_{WDICE4}$ in Fig. 9 is lower than $Loss_{DICE1}$. The main reason is that $Loss_{WDICE4}$ was computed using a weighted wrinkle map (*WWM*). As shown in Fig. 4, the *WWM* has lower values excluding the wrinkle label area as the scale value s increases. This generates relatively lower losses for false positives. In other words, $Loss_{WDICE4}$ calculates the loss with higher weights only in the actual wrinkle label area, which helps improve $Loss_{DICE1}$. But it is not appropriate to detect the final output from D_4 just because $Loss_{WDICE4}$ is lower than $Loss_{DICE1}$. The final output of wrinkle segmentation should be predicted in the final decoder D_1.

4.5. Comparison for human labeling and semi-automatic labeling

In this subsection, we compare the consistency of wrinkle labels between human and semi-automatic methods. In our test, three labeling annotators created 100 wrinkle ground truths images from 100 facial images. Next, we assumed that the 100 wrinkle ground truths were rough wrinkle labels, as shown in Fig. 2 and then generated semi-automatic wrinkle ground truths. Fig. 10 shows an example of the three ground truths created by three labeling annotators. To compare the consistency of these ground truths, we computed the correlation coefficient [48,49] as follows:

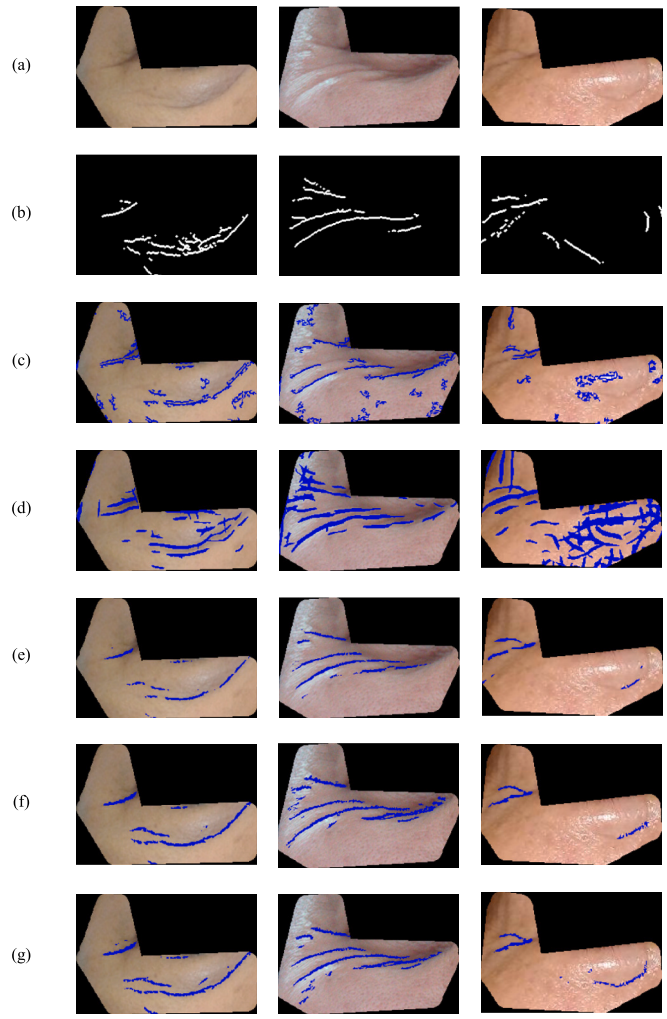

Fig. 7. Comparison of wrinkle results around the eyes. (a) is an original image, (b) is a ground truth, (c) is a result of Hessian approach [4], (d) is a result of Gabor filter approach [8], (e) is a result of U-Net without deep supervision, (f) is a result of U-Net with deep supervision, and (g) is a result of the proposed method, weighted deep supervision.

$$Correlation(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{8}$$

where X and Y are ground truths, and x_i and y_i are ith elements in X and Y, respectively. $\bar{x}$ and $\bar{y}$ represent the mean values of X and Y. As shown in Table 5, the correlation of semi-automatic labeling was higher than that of the three annotators' jobs. The semi-automatic labeling approach determines the intersection area between a rough wrinkle annotation of labeling annotators and wrinkle textures. This implies that the shape and thickness of wrinkle labels are determined by the texture map. Thus, the semi-automatic labeling approach has higher consistency.

5. Discussion

5.1. Applicability

To assess the effectiveness and generalization of weighted deep supervision, we considered retinal vessel segmentation, which deals with thin objects similar to wrinkles. We selected AG-Net [50], which uses deep supervision in this field, as the baseline and applied weighted deep supervision. AG-Net already employs attention; however, for additional comparisons, we added Atrous Spatial Pyramid Pooling (ASPP) [51] to the final encoder. To compare the performance, we selected the DRIVE dataset [52], which is widely used in vessel segmentation and consists of 20 training and 20 testing images. The resolution of each image was adjusted to 584 (H) × 584 (W) by zero-padding the horizontal axis to 584 (H) × 565 (W). Subsequently, we configured all the experimental settings to be identical to those of the wrinkle segmentation experiments. Table 6 presents the results of the deep supervision and weighted deep supervision of AG-Net. To ensure fairness, we simultaneously reported the performance of AG-Net and the results obtained in our environment. As shown in Table 6, the weighted deep supervision also demonstrated superiority in retinal vessel segmentation. Therefore, we anticipate that weighted deep supervision will help improve performance when segmenting objects similar to wrinkles or retinal vessels. The addition of ASPP enabled further performance improvements. The experimental code for retinal vessel segmentation is publicly available in our code repository, enabling easy replication.

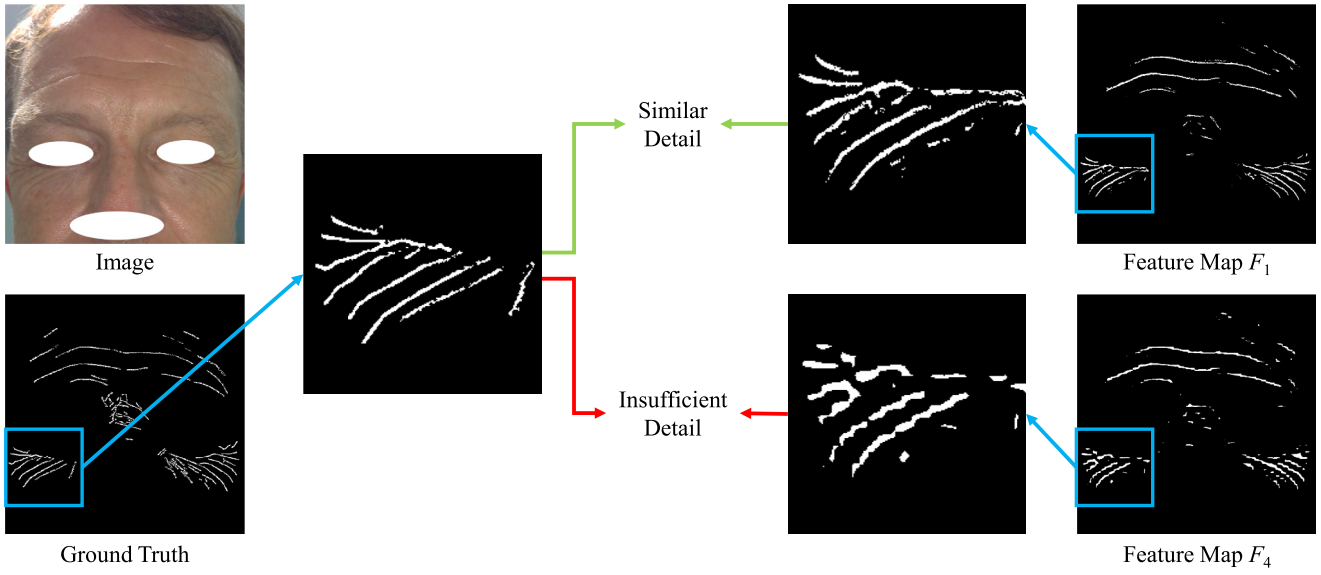

Fig. 8. Comparison of wrinkle representation between feature map F_1 and feature map F_4.

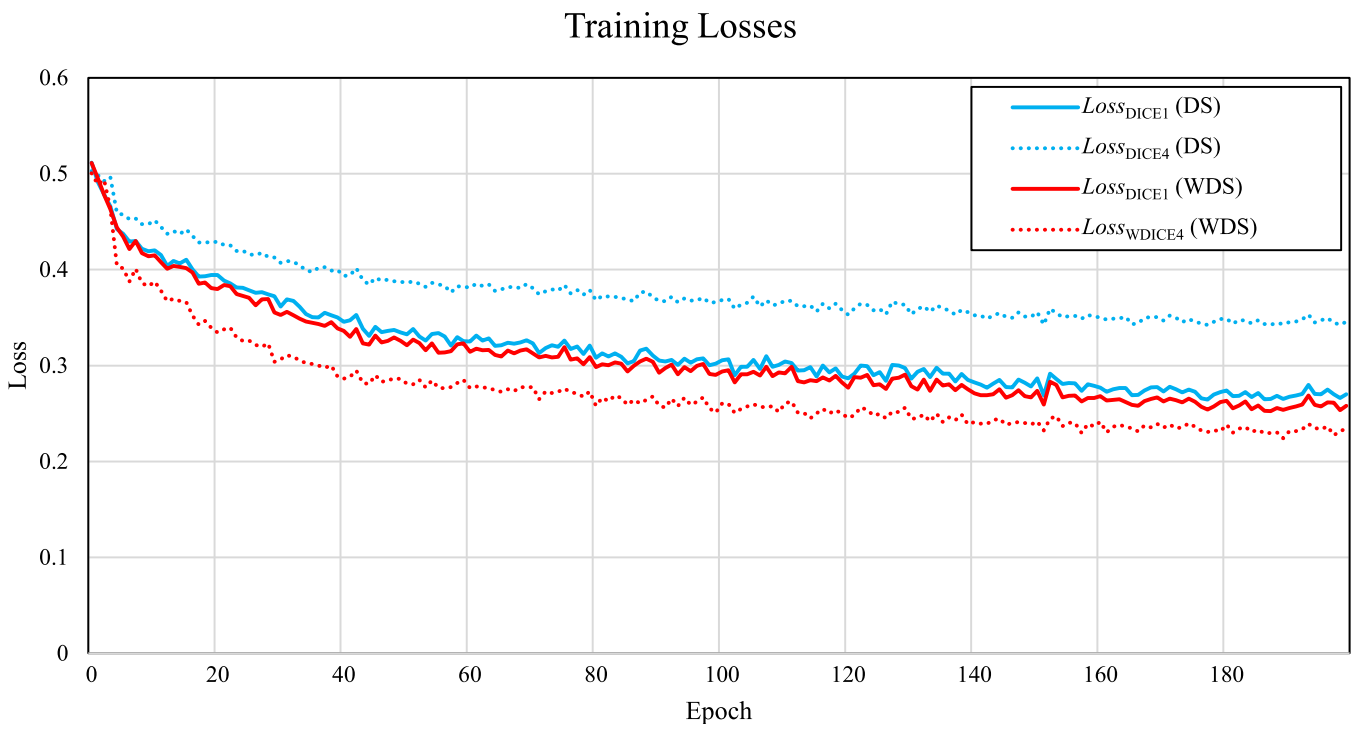

Fig. 9. Example of loss comparison between deep supervision and weighted deep supervision.

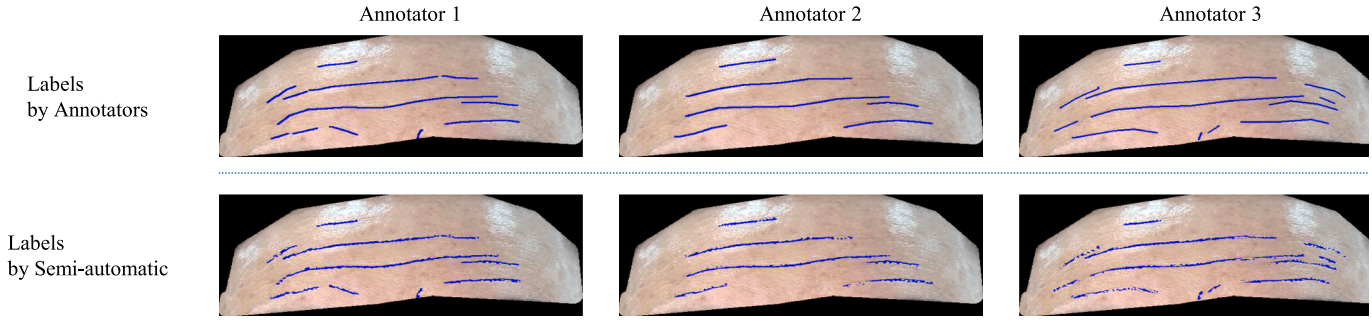

Fig. 10. An example of human labels and semi-automatic labels.

Table 5
The comparison of human and semi-automatic labels based on correlations.

	Annotator 1 and Annotator 2	Annotator 1 and Annotator 3	Annotator 2 and Annotator 3	Average
Human labels	0.378	0.308	0.259	0.315
Semi-automatic labels	0.554	0.490	0.433	0.492

Table 6
Comparison of retinal vessel segmentation performance on the DRIVE dataset using AG-Net, with weighted deep supervision (WDS) and atrous spatial pyramid pooling (ASPP).

Metric	IoU	Accuracy	Sensitivity	Specificity
AG-Net [50]	0.6965	0.9692	0.8100	0.9848
AG-Net	0.6934	0.9697	0.8081	0.9847
AG-Net + WDS	0.7036	0.9706	0.8227	0.9850
AG-Net + WDS + ASPP	0.7068	0.9709	0.8274	0.9851

5.2. Limitation

In this study, wrinkle segmentation performance was limited to 44.35 % based on the JSI. This is significantly lower than the performance levels of approximately 70–90 % in fields such as retinal vessel segmentation [50], polyp segmentation [53], and brain tumor segmentation [54]. To analyze this, we calculated the ratio of wrinkle labels in our dataset to the entire facial area using Eq. (12) as follows:

$$rate = \frac{\sum_i GT(i)}{H \times W}, \tag{12}$$

where H is the height and W is the width of GT, and $GT(i)$ is ith value of GT. Using Eq. (12), we can calculate the ratio of the wrinkle label, which results in a value of 0.015. By applying this equation to the retinal vessel dataset DRIVE, we obtained a value of 0.084. Therefore, the data imbalance was severe in our dataset because of the significantly smaller proportion of wrinkles. A case similar to ours can be observed in the research of Zheng et al. [55], where the performance of wrinkle

Table 7
Comparison of wrinkle segmentation performance with various segmentation models.

Metric	JSI	Accuracy	Sensitivity	Specificity
PSPNet	0.3537	0.9861	0.5386	0.9925
DeeplabV3+	0.3750	0.9875	0.5336	0.9935
SegNet	0.3599	0.9861	0.5553	0.9922
SegNet +WDS	0.4073	0.9867	0.5971	0.9933
U-Net (ours)	0.4297	0.9873	0.6360	0.9931
U-Net + WDS (ours)	0.4482	0.9876	0.6412	0.9937

segmentation was only 15.12 % based on the IoU criteria. Because wrinkles account for only a very small portion of the entire facial area, wrinkle segmentation is an extremely challenging problem. To improve the performance of wrinkle segmentation, it is necessary to improve U-Net, which is the baseline model of our study, or to apply other models to enhance the performance. Additionally, performing wrinkle detection by cropping the image into facial regions is expected to improve the performance, but it will also increase the time required for detection.

5.3. Comparison of segmentation model performance

In this sub-section, we aim to discuss wrinkle segmentation by comparing our proposed method with other segmentation models. We evaluated the wrinkle segmentation performance using commonly used models in general segmentation research, such as PSPNet [56], DeeplabV3+ [51], and SegNet [57], in addition to the U-Net we employed. We trained each model using one of the six folds from the training set used in Section 4. We used the same set of hyperparameters for all models. Table 7 presents the performance of each model in various metrics. PSPNet exhibited the lowest overall performance. This can be attributed to its encoder-decoder structure, which does not include skip connections, making it difficult to represent thin objects like wrinkles effectively. For DeeplabV3+, the mere application of 4× upsampling at the final output was deemed insufficient to capture the shape of wrinkles. SegNet, on the other hand, employed unpooling instead of upsampling, but it was found to be less effective for wrinkle segmentation. Similar results were observed in flood area segmentation studies using aerial photographs, indicating that SegNet performs worse than U-Net in segmenting thin and elongated objects like wrinkles. However, SegNet, like U-Net, has an encoder-decoder structure, making it easy to apply the *WDS* (weighted deep supervision) and we also observed performance improvements with 4.74 % JSI. Therefore, the U-Net architecture can be considered suitable for wrinkle detection.

6. Conclusion

In this study, we propose an improved facial wrinkle segmentation model based on weighted deep supervision. We generated weighted wrinkle maps through average pooling and upsampling, and used them to calculate more precise losses in the downsampled decoders. In our experiments, the proposed weighted deep supervision approach showed better performance than the deep supervision approach in multiple metrics. We also compared semi-automatic labels with human labels. The semi-automatic labels showed higher consistency than human labels did. However, the JSI of our proposed method was lower than that of other similar applications such as retinal vessel segmentation, polyp segmentation, brain tumor segmentation, and so on. Although the wrinkle dataset used in this study included various lighting conditions, it was determined that relying solely on improving the loss function based on the U-Net architecture led to low performance. Therefore, we plan to implement a network that is more suitable for small-object segmentation to improve wrinkle segmentation performance in future research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was result of a study on the “HPC Support” project, supported by the ‘Ministry of Science and ICT’ and NIPA.

References

[1] Aznar-Casanova J, Torro-Alves N, Fukusima S. How much older do you get when a wrinkle appears on your face? Modifying age estimates by number of wrinkles. Aging Neuropsychol Cogn 2010;17(4):406–21.
[2] Huang Y, Li Y, Fan N. Robust symbolic dual-view facial expression recognition with skin wrinkles: local versus global approach. IEEE Trans Multimed 2010;12(6):536–43.
[3] Frangi AF. Three-dimensional model-based analysis of vascular and cardiac images. Ph.D. dissertation,. Utrecht, The Netherlands: Univ. Med. Center Utrecht; 2001.
[4] Ng C-C, Yap MH, Costen N, Li B. Automatic wrinkle detection using hybrid Hessian filter. In: Computer Vision—ACCV 2014: 12th Asian conference on computer vision, Singapore, Singapore, November 1–5, 2014, revised selected papers, Part III 12; 2015. p. 609–22.
[5] Ng C-C, Yap MH, Costen N, Li B. Wrinkle detection using Hessian line tracking. IEEE Access 2015;3:1079–88.
[6] Yap MH, Alarifi J, Ng C-C, Batool N, Walker K. Automated facial wrinkles annotator. In: Proc. Eur. conf. comput. vis.; 2018. p. 676–80.
[7] Cula OG, BargoNkengne PRA, Kollias N. Assessing facial wrinkles: automatic detection and quantification. Skin Res Technol 2013;19(1):e243–51.
[8] Batool N, Chellappa R. Fast detection of facial wrinkles based on gabor features using image morphology and geometric constraints. Pattern Recognit 2015;48(3):642–58.
[9] Qin X, Ban Y, Wu P, Yang B, Liu S, Yin L, et al. Improved image fusion method based on sparse decomposition. Electronics 2022;11(15):2321. https://doi.org/10.3390/electronics11152321.
[10] Liu H, Yue Y, Liu C, Spencer BF, Cui J. Automatic recognition and localization of underground pipelines in GPR B-scans using a deep learning model. Tunnell Undergr Space Technol 2023;134:104861. https://doi.org/10.1016/j.tust.2022.104861.
[11] Lu S, Ding Y, Liu M, Yin Z, Yin L. Multiscale feature extraction and fusion of image and text in VQA. Int J Comput Intell Syst 2023;16(1):54. https://doi.org/10.1007/s44196-023-00233-6.
[12] Liu H, Liu M, Li D, Zheng W, Yin L. Recent advances in pulse-coupled neural networks with applications in image processing. Electronics 2022;11(20). https://doi.org/10.3390/electronics11203264.
[13] Mamalakis M, Garg P, Nelson T, Lee J, Swift AJ, Wild JM, et al. Artificial intelligence framework with traditional computer vision and deep learning approaches for optimal automatic segmentation of left ventricle with scar. Artif Intell Med 2023;102610.
[14] Hao D, Ahsan M, Salim T, Duarte-Rojo A, Esmaeel D, Zhang Y, et al. A self-training teacher-student model with an automatic label grader for abdominal skeletal muscle segmentation. Artif Intell Med 2022;132:102366.
[15] Sun Y, Yang H, Zhou J, Wang Y. ISSMF: integrated semantic and spatial information of multi-level features for automatic segmentation in prenatal ultrasound images. Artif Intell Med 2022;125:102254.
[16] Zhang Y, Jiao R, Liao Q, Li D, Zhang J. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artif Intell Med 2023;138:102476.
[17] Gruber N, Galijasevic M, Regodic M, Grams AE, Siedentopf C, Steiger R, et al. A deep learning pipeline for the automated segmentation of posterior limb of internal capsule in preterm neonates. Artif Intell Med 2022;132:102384.
[18] Sáenz-Gamboa JJ, Domenech J, Alonso-Manjarrés A, Gómez JA, de la Iglesia-Vayá M. Automatic semantic segmentation of the lumbar spine: clinical applicability in a multi-parametric and multi-center study on magnetic resonance images. Artif Intell Med 2023;140:102559.
[19] Khan RA, Luo Y, Wu F-X. RMS-U-Net: residual multi-scale U-Net for liver and lesion segmentation. Artif Intell Med 2022;124:102231.
[20] Boutillon A, Borotikar B, Burdin V, Conze P-H. Multi-structure bone segmentation in pediatric MR images with combined regularization from shape priors and adversarial network. Artif Intell Med 2022;132:102364.
[21] Mecheter I, Abbod M, Amira A, Zaidi H. Deep learning with multiresolution handcrafted features for brain MRI segmentation. Artif Intell Med 2022;131:102365.
[22] Chai C, Qiao P, Zhao B, Wang H, Liu G, Wu H, et al. Brain gray matter nuclei segmentation on quantitative susceptibility mapping using dual-branch convolutional neural network. Artif Intell Med 2022;125:102255.
[23] Xiong H, Liu S, Sharan RV, Coiera E, Berkovsky S. Weak label based Bayesian U-Net for optic disc segmentation in fundus images. Artif Intell Med 2022;126:102261.

[24] Jin K, Yan Y, Wang S, Yang C, Chen M, Liu X, et al. iERM: an interpretable deep learning system to classify epiretinal membrane for different optical coherence tomography devices: a multi-center analysis. J Clin Med 2023;12(2). https://doi.org/10.3390/jcm12020400.
[25] Jin K, Gao Z, Jiang X, Wang Y, Ma X, Li Y, et al. MSHF: a multi-source heterogeneous fundus (MSHF) dataset for image quality assessment. Sci Data 2023;10(1):286. https://doi.org/10.1038/s41597-023-02188-x.
[26] Gao Z, Pan X, Shao J, Jiang X, Su Z, Jin K, et al. Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning. Br J Ophthal 2022;2022-321472. https://doi.org/10.1136/bjo-2022-321472.
[27] Ben Hamida A, Devanne M, Weber J, Truntzer C, Derangère V, Ghiringhelli F, et al. Weakly supervised learning using attention gates for colon cancer histopathological image segmentation. Artif Intell Med 2022;133:102407.
[28] Ao J, Shao X, Liu Z, Liu Q, Xia J, Shi Y, et al. Stimulated Raman scattering microscopy enables Gleason scoring of prostate core needle biopsy by a convolutional neural network. Cancer Res 2023;83(4):641–51. https://doi.org/10.1158/0008-5472.CAN-22-2146.
[29] Kim S, Lee C, Jung G, Yoon H, Lee J, Yoo S. Facial acne segmentation based on deep learning with center point loss. In: 2023 IEEE 36th international symposium on computer-based medical systems (CBMS), L'Aquila, Italy; 2023. p. 678–83.
[30] Yoon H, Kim S, Lee J, Yoo S. Deep-learning-based morphological feature segmentation for facial skin image analysis. Diagnostics 2023;13(11):1894. Published 2023 May 29, https://doi.org/10.3390/diagnostics13111894.
[31] Jung G, Kim S, Lee J, Yoo S. Deep learning-based optical approach for skin analysis of melanin and hemoglobin distribution. J Biomed Opt 2023;28:035001.
[32] Lee C, Yoo S, Kim S, Lee J. Progressive weighted self-training ensemble for multi-type skin lesion semantic segmentation. IEEE Access 2022;10:132376–83. https://doi.org/10.1109/ACCESS.2022.3222788.
[33] Kim S, Yoon H, Lee J, Yoo S. Semi-automatic labeling and training strategy for deep learning-based facial wrinkle detection. In: 2022 IEEE 35th international symposium on computer-based medical systems (CBMS); 2022. p. 383–8.
[34] Bradley D, Roth G. Adaptive thresholding using the integral image. J Graph Tools 2007;12(2):13–21.
[35] Fu H, Cheng J, Xu Y, Wong DWK, Liu J, Cao X. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans Med Imaging 2018;37(7):1597–605.
[36] Zhang S, Fu H, Yan Y, Zhang Y, Wu Q, Yang M, et al. Attention guided network for retinal image segmentation. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2019. p. 797–805.
[37] Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, et al. U-Net 3+: a full-scale connected U-Net for medical image segmentation. In: ICASSP 2020—2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2020. p. 1055–9.
[38] Lumini. KIOSK v2. https://en.lulu-lab.com/lumini.html.
[39] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2015. p. 234–41.
[40] Github. https://github.com/milesial/Pytorch-U-Net.
[41] Jana R, Datta D, Saha R. Age estimation from face image using wrinkle features. Proc Comput Sci 2015;46:1754–61.
[42] H. Razalli, R. W. O. K. Rahmat, F. Khalid, and P. S. Sulaiman, 'Age range estimation based on facial wrinkle analysis using hessian based filter', in Advanced computer and communication engineering technology: proceedings of ICOCOE 2015, 2016, pp. 759–769.
[43] Batool N, Chellappa R. Modeling and detection of wrinkles in aging human faces using marked point processes. In: Computer vision—ECCV 2012. Workshops and demonstrations: Florence, Italy, October 7–13, 2012, proceedings, part II. vol. 12; 2012. p. 178–88.
[44] Jin H, Liao S, Shao L. Pixel-in-pixel net: towards efficient facial landmark detection in the wild. Int J Comput Vis 2021;129:3174–94.
[45] Loshchilov I, Hutter F. Sgdr: stochastic gradient descent with warm restarts. In: 5th International Conference on Learning Representations (ICLR); 2017.
[46] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: 3th International Conference on Learning Representations (ICLR); 2015.
[47] Ranjbarzadeh R, Caputo A, Tirkolaee EB, Ghoushchi SJ, Bendechache M. Brain tumor segmentation of MRI images: a comprehensive review on the application of artificial intelligence tools. Comput Biol Med 2023;152:106405.
[48] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.
[49] Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology 2003;227(3):617–28.
[50] Zhang S, et al. Attention guided network for retinal image segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part I. vol. 22; 2019. p. 797–805.
[51] Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV; 2018.
[52] Staal J, Abràmoff MD, Niemeijer M, Viergever MA, Van Ginneken B. Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging 2004;23(4):501–9.
[53] Fan D-P, et al. Pranet: parallel reverse attention network for polyp segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, proceedings, part VI. 23; 2020. p. 263–73.
[54] Ranjbarzadeh R, Bagherian Kasgari A, Jafarzadeh Ghoushchi S, Anari S, Naseri M, Bendechache M. Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. Sci Rep 2021;11(1):1–17.
[55] Zheng Q, Purwar A, Zhao H, Lim GL, Li L, Behera D, et al. Automatic facial skin feature detection for everyone. In: Proc. IS&T int'l. symp. on electronic imaging: imaging and multimedia analytics at the edge; 2022. 300-1–300-6.
[56] Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA; 2017. p. 6230–9. https://doi.org/10.1109/CVPR.2017.660.
[57] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39(12):2481–95. 1 Dec. https://doi.org/10.1109/TPAMI.2016.2644615.