

Facial Wrinkle Segmentation for Cosmetic Dermatology: Pretraining with Texture Map-Based Weak Supervision

Junho Moon^{1[0009-0004-3522-6357]}, Haejun Chung^{*1[0000-0001-8959-237X]}, and Ikbeom Jang^{*2[0000-0002-6901-983X]}

¹ Hanyang University, Seoul 04763, Republic of Korea
 {jhmoon6807, haejun}@hanyang.ac.kr

² Hankuk University of Foreign Studies, Yongin 17035, Republic of Korea
 ijang@hufs.ac.kr

Abstract. Facial wrinkle detection plays a crucial role in cosmetic dermatology. Precise manual segmentation of facial wrinkles is challenging and time-consuming, with inherent subjectivity leading to inconsistent results among graders. To address this issue, we propose two solutions. First, we build and release the first public facial wrinkle dataset, ‘FFHQ-Wrinkle’, an extension of the NVIDIA FFHQ dataset. It includes 1,000 images with human labels and 50,000 images with automatically generated weak labels. This dataset could serve as a foundation for the research community to develop advanced wrinkle detection algorithms. Second, we introduce a simple training strategy utilizing texture maps, applicable to various segmentation models, to detect wrinkles across the face. Our two-stage training strategy first pre-train models on a large dataset with weak labels ($N=50k$), or masked texture maps generated through computer vision techniques, without human intervention. We then finetune the models using human-labeled data ($N=1k$), which consists of manually labeled wrinkle masks. The network takes as input a combination of RGB and masked texture map of the image, comprising four channels, in finetuning. We effectively combine labels from multiple annotators to minimize subjectivity in manual labeling. Our strategies demonstrate improved segmentation performance in facial wrinkle segmentation both quantitatively and visually compared to existing pretraining methods. The dataset is available at <https://github.com/labhai/ffhq-wrinkle-dataset>.

Keywords: Facial wrinkle segmentation · Weakly supervised learning · Texture map pretraining · Transfer learning

1 Introduction

With the growing interest in dermatological diseases and skin aesthetics, predicting facial wrinkles is becoming increasingly significant. Facial wrinkles serve

*Corresponding authors

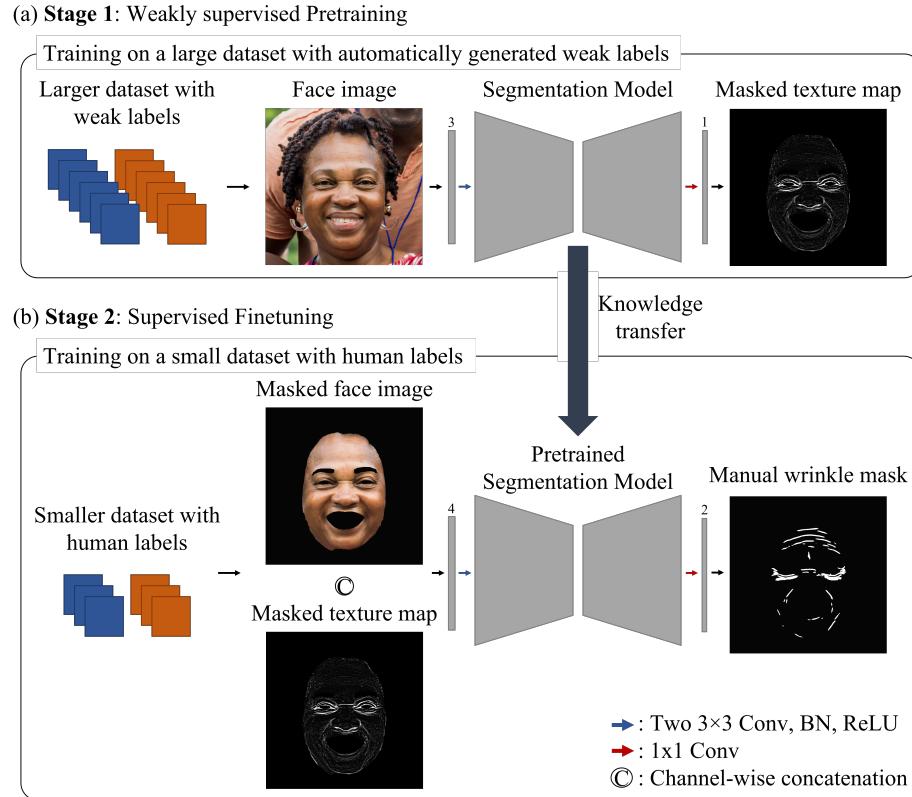


Fig. 1. Two-stage training for facial wrinkle segmentation. (a) Weakly supervised pre-training stage: the model learns to extract masked texture maps from RGB face images. (b) Supervised finetuning stage: the model refines its ability to extract facial wrinkles from RGB-masked face images and masked texture maps. The model parameters are initialized with the weights from the weakly supervised pretraining stage.

as critical indicators of aging [2,19,20], and are essential for evaluating skin conditions [29,13], diagnosing dermatological disorders [30], and planning pre-treatment protocols for skin management [1,27]. Nevertheless, the manual detection of facial wrinkles poses considerable challenges. Accurate detection and analysis of facial wrinkles necessitate a high level of expertise, typically available only through well-trained professionals such as dermatologists. This process is time-consuming and entails substantial costs due to the extensive time and effort required by the experts.

Recently, numerous studies have focused on the automatic segmentation of facial wrinkles through the application of deep learning techniques [25,26,14,15,4,34]. Nevertheless, these deep learning-based approaches are notably data-intensive. Due to the intricate distribution of facial wrinkles across the face, analyzing ex-

tensive collections of images can be exceedingly resource-intensive if each wrinkle must be individually evaluated. Furthermore, the manual analysis procedure is fraught with subjectivity. The assessments of individual experts can differ significantly based on their experience, level of training, and personal biases, thereby complicating the consistency and reproducibility of the analysis results.

To address these challenges, we propose a two-stage training strategy, as illustrated in Fig. 1. This approach utilizes computer vision techniques, specifically filters, to generate many weakly labeled wrinkle masks ($N=50,000$) without human intervention for weakly supervised pretraining. A smaller set of accurately labeled wrinkle masks ($N=1,000$) is employed for supervised finetuning. This method significantly decreases the time and cost associated with manual wrinkle labeling, providing substantial advantages over traditional methodologies. To ensure the development of a generalized and robust model, we conducted experiments using a dataset comprising images captured from various angles, lighting conditions, races, ages, and skin conditions. We quantitatively analyzed the challenges associated with consistent manual wrinkle labeling across such a diverse dataset and integrated data labeled by multiple annotators to reduce subjectivity during the finetuning stage. No public dataset exists for full-face wrinkle segmentation, although there are a few private datasets. To address this gap, we have made our dataset publicly accessible to enhance the reproducibility and reliability of our results. This initiative aims to reduce the manual labeling costs for future research and serve as a benchmark dataset.

2 Related works

2.1 Deep learning-based facial wrinkle segmentation

Deep learning-based methods for facial wrinkle segmentation aim to enable neural network models to learn the features necessary for accurate wrinkle detection autonomously. Kim et al. [14] introduced a semi-automatic labeling strategy to enhance performance by extracting texture maps from face images and combining them with roughly labeled wrinkle masks, utilizing a U-Net architecture [23] for segmentation. In a subsequent study [15], they further improved segmentation accuracy by implementing a weighted deep supervision technique, which employs a weighted wrinkle map to more precisely calculate the loss for the downsampled decoder, outperforming traditional deep supervision methods. Yang et al. [34] developed Striped WriNet, which integrates a Striped Attention Module composed of Multi-Scale Striped Attention and Global Striped Attention within a U-shaped network. This approach applies an attention mechanism across multiple scales, effectively segmenting both coarse and fine wrinkles.

2.2 Weakly supervised learning

Weakly supervised learning is a methodology that trains models using incomplete or inaccurate labeled data instead of fully labeled data in situations where strong

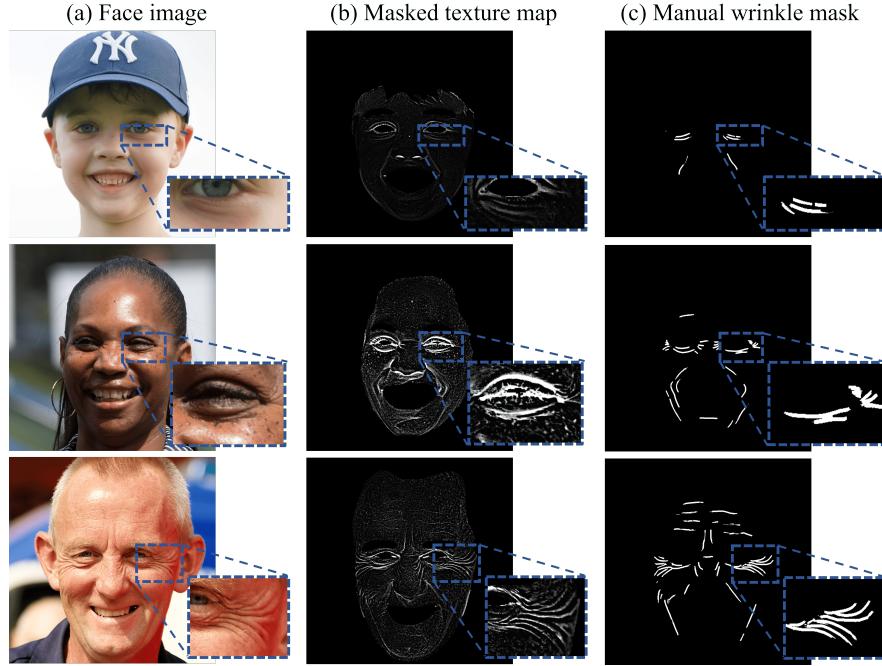


Fig. 2. Training Dataset. (a) High-resolution face images. (b) Masked texture maps extracted from face images, which include information about facial features. (c) Reliable manual wrinkle masks created by combining the results of multiple annotators.

supervision information is lacking [36]. Xu et al. [33] proposed CAMEL, a weakly supervised learning framework that uses a MIL-based label expansion technique to divide images into grid-shaped instances and automatically generate instance-level labels, enabling histopathology image segmentation with only image-level labels. Shen et al. [11] trained a deep learning model using only scribbles on whole tumors and healthy brain tissue, along with global labels for the presence of each substructure, to segment all sub-regions of brain tumors.

3 Dataset

3.1 Dataset specifications

The first public facial wrinkle dataset, ‘FFHQ-Wrinkle’, comprises pairs of face images and their corresponding wrinkle masks. We focused on wrinkle labels while utilizing the existing face image dataset FFHQ (Flickr-Faces-HQ) [12], which contains 70,000 high-resolution (1024x1024) face images captured under various angles and lighting conditions. The dataset we provide consists of one set of manually labeled wrinkle masks ($N=1,000$) and one set of ‘weak’ wrinkle masks, or masked texture maps, generated without human labor ($N=50,000$).

Table 1. Demographic attributes of the dataset. The ‘Human-labeled’ data represents the 1,000 face images manually labeled by human annotators and the ‘Weakly-labeled’ data refers to the 50,000 images labeled without human intervention.

Dataset		Human-labeled	Weakly-labeled
Sample size		1000	50000
Age	0-9 / 10-19 / 20-29 / 30-39 / 40-49 / 50-69 / 70+	66 / 68 / 233 / 246 / 186 / 161 / 40	7030 / 4448 / 13804 / 10960 / 6931 / 5550 / 1277
Sex	Male / Female	471 / 529	26929 / 23071
Race/ Ethnicity	White / Asian / Latino Hispanic / Black / Middle Eastern / Indian	587 / 210 / 67 / 81 / 37 / 18 /	29728 / 11121 / 3895 / 2383 / 2053 / 820

We selected 50,000 images from the FFHQ dataset, specifically image IDs 00000 to 49999. We used these 50,000 face images to create the weakly labeled wrinkles and randomly sampled 1,000 images from these to create the ground truth wrinkles. The methods for generating weakly labeled wrinkles and ground truth wrinkles are discussed in Section 4.2. Table 1 summarizes estimated demographic information of the dataset—i.e. age, race, and sex. The age and sex data were sourced from the FFHQ-Aging [22] dataset, where at least three annotators labeled each image. The race/ethnicity attribute was obtained through facial attribute analysis using the DeepFace[‡] framework. Hence, the demographic information may include errors. As illustrated in Fig. 2, the dataset consists of individuals of varying ages, sex, and race/ethnicity, featuring a range of skin conditions such as freckles, acne, and pigmentation. This diversity makes the dataset particularly suitable for training models to handle the wide array of skin conditions encountered in clinical settings. The dataset is publicly available at <https://github.com/labhai/fhq-wrinkle-dataset>.

3.2 Ground truth wrinkle annotation

For ground truth wrinkles, we manually annotated the face images. The annotation process involved three annotators with extensive experience in image processing and analysis. Wrinkles can be categorized into two types—dynamic wrinkles and static wrinkles [31]. Dynamic wrinkles are formed by facial muscles and appear with expressions but disappear when the face is at rest. Static (permanent) wrinkles are visible even when the face is at rest and result from the repeated formation of dynamic wrinkles over time. We annotated both types of wrinkles without distinguishing between them. Given the subjectivity inherent in wrinkle data, a consistent standard for wrinkle assessment was established prior to the commencement of labeling. The annotators conducted three synchronization sessions to minimize inter-rater variability. The annotation primarily targeted the forehead, crow’s feet, and nasolabial folds, encompassing the overall facial area. Due to the high resolution and diversity of the dataset—comprising various races, skin conditions, backgrounds, and angles—achieving consistent labeling results proved challenging, even with established standards for wrinkle

[‡]<https://github.com/serengil/deepface>



Fig. 3. Ambiguity in wrinkle evaluation. The labeling results from three annotators for the same image are different.

Table 2. Inter-rater agreement of manual wrinkle annotation. The Jaccard similarity index and Pearson correlation coefficient between different annotators are analyzed.

Metric	Annotators A&B	Annotators B&C	Annotators A&C	Average
Jaccard similarity index	0.2631	0.2962	0.3182	0.2925
Pearson correlation coefficient	0.4167	0.4559	0.4928	0.4551

assessment, as illustrated in Fig. 3. Consequently, as demonstrated in Table 2, the inter-rater agreement was low, underscoring the highly subjective nature of wrinkle assessments.

4 Method

4.1 Model architecture

We evaluated our proposed method using the U-Net [23] and Swin UNETR [9] architectures, with U-Net serving as the base model for ablation studies and additional experiments. As depicted in Fig. 1, the U-Net model features a standard architecture comprising four encoder blocks and four decoder blocks. The Swin UNETR model employs an encoder with a window size of 16 and patches of size 4x4, projecting the input patch into a 48-dimensional embedding space. This model includes four encoder blocks, each consisting of two successive Swin Transformer blocks [16], and four decoder blocks.

4.2 Training strategy

We train the segmentation model using a substantial number of masked texture maps in a weakly supervised manner, followed by finetuning with a smaller set of reliably manually labeled wrinkle masks in a supervised manner. This training strategy, which involves finetuning the weights of a pretrained model that extracts facial textures using human-labeled wrinkle data, significantly enhances the model’s capability to detect facial wrinkles. The overall training pipeline is illustrated in Fig. 1.

Weakly supervised pretraining stage In the pretraining stage, we utilized weakly labeled wrinkle data automatically extracted through computer vision techniques without human intervention as the ground truth. Fig. 4 illustrates the pipeline for generating weakly labeled wrinkles for the weakly supervised pretraining stage. Utilizing Equation (1), we extracted the texture map [14] from the face image through a Gaussian kernel-based filter.

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

where G represents the Gaussian kernel, σ denotes its standard deviation, $I_{G(\sigma)}$ is the Gaussian filtered image, and (x, y) are the pixel coordinates in the image. Following the methodology in [14], we set the Gaussian kernel's standard deviation to 5 and its size to 21x21 for texture map extraction. The extracted texture map contains detailed information about the contours, curves, and skin textures of the face image. However, as the texture map includes numerous false positives from the background, we employ a BiSeNet [35] architecture-based facial parsing deep learning model[§] to mask non-facial regions, resulting in the final masked texture map used as ground truth. We avoid converting the masked texture map into a binary mask due to the variability in the size, shape, and depth of wrinkles, which makes determining an appropriate threshold challenging. Fig. 2-(b) shows the masked texture map used as the final ground truth in the weakly supervised pretraining stage.

In the weakly supervised pretraining stage, the model takes a 3-channel RGB face image as input and outputs a 1-channel masked texture map (Fig. 1-(a)). We use mean squared error (MSE) loss [21] to optimize the model, calculated as shown in equation (2).

$$MSE(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (2)$$

where \hat{y}_i and y_i are the model output and the masked texture map, respectively.

Supervised finetuning stage For the ground truth in the finetuning stage, we utilized human-labeled wrinkle data generated as described in Section 3.2. Fig. 5 illustrates the pipeline of the ground truth generation of the wrinkle mask. To produce a reliable ground truth wrinkle mask, we used majority voting to retain only the pixels that were labeled by at least two groups, thereby reducing variability among the annotators. Fig. 2-(c) displays the manual wrinkle mask used as the final ground truth in the supervised finetuning stage. As model inputs, we use masked face images, where non-facial regions were masked using a facial-parsing model. Additionally, we included masked texture maps, which were used as ground truth in the pretraining stage, as auxiliary inputs.

In the supervised finetuning stage, the model takes as input a 3-channel RGB face image with only the facial regions and a 1-channel masked texture map.

[§]<https://github.com/zllrunning/face-parsing.PyTorch>

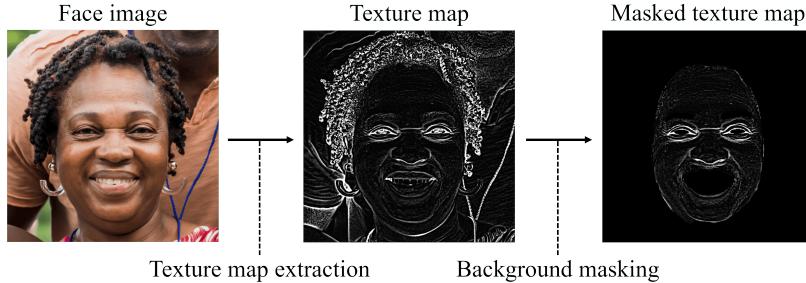


Fig. 4. Weakly labeled wrinkle generation pipeline. After extracting the texture map from the face image, we mask the non-facial regions to generate a masked texture map containing information on facial features. This masked texture map is then used as a weakly labeled wrinkle.

It then produces a 2-channel output indicating the presence of wrinkles and background. This stage begins with the model parameters from the pretraining stage, where the model was weakly supervised to extract masked texture maps from face images. Using transfer learning, we refine the model by adjusting its weights with manually labeled wrinkle masks. This process enhances the model’s ability to detect facial wrinkles by building on the general facial texture extraction skills developed during pretraining. We optimize the model using soft Dice loss [5], as shown in equation (3).

$$DL(p, g) = 1 - \frac{1}{C} \sum_{c=1}^C \frac{2 \sum_{i=1}^N p_{i,c} g_{i,c}}{\sum_{i=1}^N p_{i,c} + \sum_{i=1}^N g_{i,c}} \quad (3)$$

where C is the total number of classes, N is the total number of pixels, $p_{i,c}$ represents the predicted probability for pixel i belonging to class c , and $g_{i,c}$ represents the ground truth label for pixel i belonging to class c , respectively.

5 Experiments

5.1 Implementation details

In both the weakly supervised pretraining and supervised finetuning stages, we utilize the original 1024x1024 image-label pairs as inputs without resizing. The AdamW optimizer [18] is employed, configured with a weight decay of 0.05, β_1 set to 0.9, and β_2 set to 0.999. We also implement the SGDR scheduler [17]. To maintain dataset diversity, we randomly apply various augmentations, including horizontal flipping, scaling, affine transformation, elastic transformation, grid distortion, and optical distortion during training. The dataset is partitioned into 80% for training, 10% for validation, and 10% for testing.

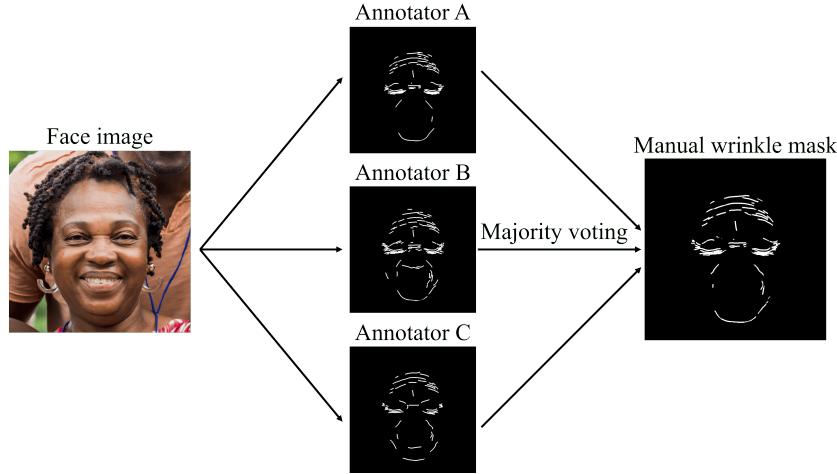


Fig. 5. Ground truth wrinkle generation pipeline. We combine data labeled by multiple annotators through majority voting to create a reliable ground truth wrinkle.

Weakly supervised pretraining stage In the weakly supervised pretraining stage, the model is trained for 300 epochs. The SGDR scheduler starts with an initial period of 100 epochs, with the learning rate beginning at a maximum of 0.001 and decaying to 0 over the period. At the end of each period, the length of the next period doubles that of the previous one. The batch size is 26 for U-Net and 22 for Swin UNETR. All pretraining processes were performed on an NVIDIA A100 Tensor Core GPU.

Supervised finetuning stage In the supervised finetuning stage, the U-Net model is finetuned for 150 epochs, while the Swin UNETR model is finetuned for 300 epochs. The batch size is 14 for both models. The SGDR scheduler’s initial period length is set to 50 epochs for U-Net and 100 epochs for Swin UNETR. The learning rate starts at a maximum of 0.0001 and decreases to 0 within each period. At the end of each period, the length of the next period doubles that of the previous one, with the maximum learning rate set to 90% of the last period’s maximum. All finetuning processes are performed on RTX A6000 and RTX 6000 Ada GPUs.

5.2 Evaluation metrics

To evaluate the performance of the final finetuned model in wrinkle segmentation, we use the Jaccard Similarity Index (JSI), F1-score, and Accuracy (Acc).

The Jaccard Similarity Index measures the overlap between the predicted wrinkle regions and the ground truth regions, defined as follows:

$$\text{JSI} = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

where A is the predicted segmentation, and B is the actual label.

The F1-score is the harmonic mean of precision and recall, while accuracy measures the proportion of correctly predicted pixels out of the total pixels. They are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

where TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives, and TN is the number of true negatives.

5.3 Results

To evaluate the performance of our proposed method, we first compare it with the latest methods: the semi-automatic labeling and weighted deep supervision method [15], and the Striped WriNet method [34]. Because the primary contribution of this work is the pretraining strategy, we also compare it with other pretraining techniques. They include using ImageNet pretrained models and self-supervised learning methods. For the ImageNet pretrained models, we replace the encoder part of the U-shape architecture with models pretrained on the ImageNet-1K dataset [24]; specifically, we use ResNet-50 [10] for U-Net and Swin-T [16] for Swin UNETR. For the self-supervised learning methods, we use denoising self-supervised learning [3] for pretraining U-Net, setting the Gaussian distribution’s standard deviation to 0.2, and masked image prediction [32] for pretraining Swin UNETR, using 32x32 masked patches and a 60% masking ratio. All training hyperparameters follow those specified in Section 5.1. To assess performance in scenarios with very limited labeled data, we train our model on the full training set (100%, N=800) and on a randomly sampled subset (5%, N=40).

The proposed method outperforms the latest wrinkle segmentation methods and the ones using the same model architectures with different pertaining methods. The performance gap is much larger in data-limited situations—i.e., fine-tuned on 5% of the manually-labeled data. Table 3 shows quantitative comparisons of wrinkle segmentation performance for each method using U-Net and Swin UNETR architectures. Our method consistently achieves the highest performance across both datasets and architectures. Fig. 6 presents a qualitative comparison of our method with denoising pretraining using U-Net, which is the next best performing method in experiments using 100% of the data.

Table 3. Quantitative comparisons of facial wrinkle segmentation performance. Our method is compared against two latest wrinkle segmentation methods, models trained without pretraining, and models using different pretraining strategies. These pretraining techniques include masked image prediction, denoising, and pretraining encoders using the ImageNet-1K dataset.

Method		100% (N=800)			5% (N=40)			n_{params}
		JSI	F1-score	Acc	JSI	F1-score	Acc	
Semi automatic labeling + WDS [15]		0.4552	0.6256	0.9954	0.3384	0.5057	0.9928	17.269M
Striped WriNet [34]		0.4665	0.6294	0.9956	0.2382	0.3761	0.9903	6.223M
Swin UNETR with pretraining	No pretraining	0.4220	0.5858	0.9949	0.2545	0.3944	0.9932	25.153M
	ImageNet-1K [24] (Swin-T [16])	0.4385	0.6028	0.9952	0.2877	0.4351	0.9939	100.56M
	Masked image modeling [32]	0.4450	0.6079	0.9954	0.2963	0.4452	0.9937	25.153M
	Texture map (ours)	0.4643	0.6271	0.9953	0.3416	0.4970	0.9944	25.155M
U-Net with pretraining	No pretraining	0.4638	0.6278	0.9955	0.3021	0.4551	0.9918	17.263M
	ImageNet-1K [24] (ResNet-50 [10])	0.4664	0.6296	0.9955	0.3428	0.5018	0.9934	32.521M
	Denoising [3]	0.4709	0.6339	0.9955	0.2840	0.4338	0.9898	17.263M
	Texture map (ours)	0.4831	0.6442	0.9957	0.3512	0.5116	0.9929	17.264M

Table 4. Ablation study of the effectiveness of adding a masked texture map as an additional model input. We conduct experiments using U-Net. The segmentation performance improves when using the masked texture map as an additional input during finetuning after texture map training.

Method	Model input	100% (N=800)			5% (N=40)			n_{params}
		JSI	F1-score	Acc	JSI	F1-score	Acc	
No pretraining	RGB (3-ch)	0.4638	0.6278	0.9955	0.3021	0.4551	0.9918	17.263M
	RGB+Texture (4-ch)	0.4606	0.6221	0.9954	0.3208	0.4743	0.9924	
Texture map pretraining	RGB (3-ch)	0.4796	0.6422	0.9957	0.3442	0.5051	0.9919	17.264M
	RGB+Texture (4-ch, ours)	0.4831	0.6442	0.9957	0.3512	0.5116	0.9929	

5.4 Ablation study

Incorporating the masked texture map as an additional input during the finetuning stage led to significant improvements in wrinkle segmentation, demonstrating the effectiveness of our approach. Table 4 presents quantitative comparisons using the U-Net architecture to assess the benefits of including a 1-channel masked texture map as an additional input during finetuning. We compare our pretraining method (Texture map pretraining) with a conventional approach (No pretraining), which is trained solely on manually labeled data, both with (RGB+Texture) and without (RGB) the additional masked texture map input.

6 Discussion

Our approach achieves state-of-the-art performance when compared to two publicly released models specifically designed for wrinkle segmentation, in addition

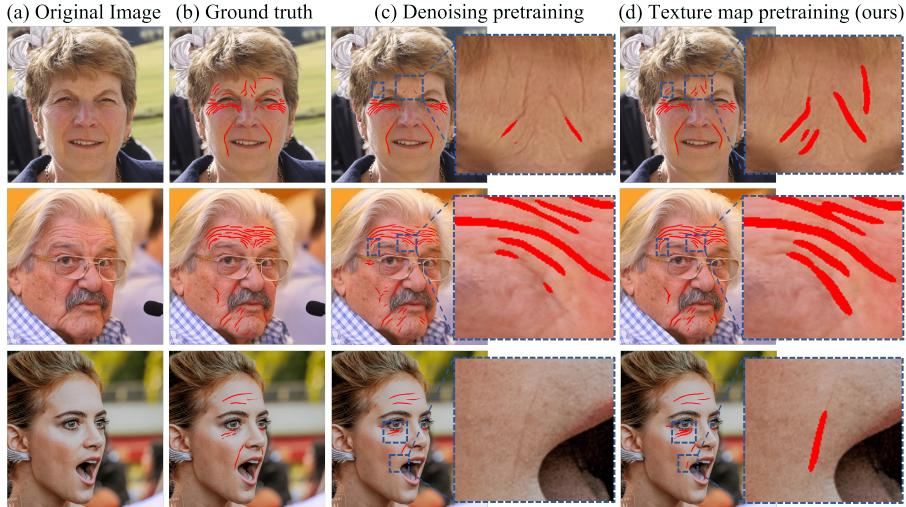


Fig. 6. Qualitative comparison against the denoising pretraining method. The blue boxes highlight areas with significant visual differences. (a) Face image. (b) Ground truth wrinkle. (c) Predicted wrinkles from a model using self-supervised learning with denoising pretraining, followed by finetuning with a manual wrinkle mask. (d) Predicted wrinkles from our model, trained with weak supervision using a masked texture map and then finetuned with a manual wrinkle mask.

to outperforming ImageNet pretrained models and self-supervised learning methods. We demonstrate that our two-stage training strategy significantly enhances wrinkle segmentation efficiency. Furthermore, our approach shows the potential to achieve high performance with limited data, which could enhance scalability and flexibility in clinical settings. By using a large amount of weakly labeled data obtained automatically through filters for weakly supervised training and then finetuning with a small amount of reliable manually labeled data, we significantly reduce the time and cost required for manual labeling while improving the segmentation performance of facial wrinkles. To minimize subjectivity in the manual labeling process, we effectively combine data labeled by multiple annotators, resulting in more reliable training data. Additionally, to enhance the reproducibility of our research and reduce the manual labeling costs for subsequent studies, we release the dataset publicly available, which can also serve as a benchmark dataset for future research. The performance improvement of facial wrinkle segmentation through transfer learning has not been conducted in previous research, indicating that our approach can be efficiently integrated into various tasks related to facial wrinkle detection and segmentation tasks. Additionally, since this research falls under the broader category of thin object detection tasks, it is expected to be widely applicable to studies requiring segmentation of thin objects (e.g., fundus imaging, vascular imaging).

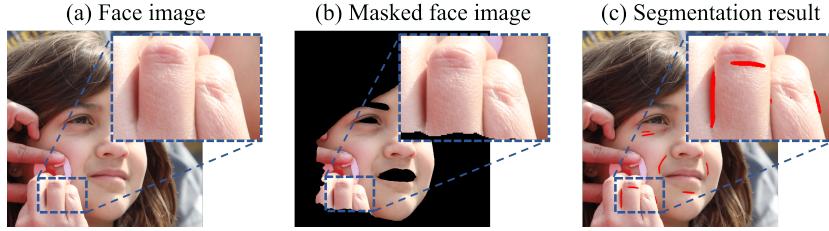


Fig. 7. Example of a false wrinkle detection. (a) Face image. (b) Masked face image used as the model input during the finetuning stage. (c) Visualization of the model’s predicted segmentation after the finetuning stage.

According to our experimental results, the performance of the Swin UNETR, a hybrid transformer-CNN architecture, is lower compared to the standard CNN-based U-Net. In our case, the dataset used for finetuning is relatively small, making it insufficient to generalize transformer models, which primarily perform well in data-intensive environments due to their low inductive bias [6]. Especially in the case of wrinkles, the relationship between adjacent pixels (skin) plays a crucial role in their assessment. Therefore, the CNN-based standard U-Net, which excels at capturing local information, tends to outperform the Swin UNETR, which includes transformer blocks specialized in capturing global context through multi-head attention mechanisms. Nevertheless, our experimental results show that the performance of Swin UNETR progressively improves through our method, suggesting that with more data and longer pretraining, there is significant potential for performance enhancement. Note that accuracy is very high in all experiments since wrinkles occupy a very small proportion of the face and most of the predictions are background pixels.

However, our approach has limitations. As shown in Fig. 7, objects similar to wrinkles, such as hair or fingers covering the face, are mistakenly recognized as wrinkles in the images. This results in false positives during the wrinkle segmentation process. To address this issue, upcoming studies will focus on developing techniques that can accurately segment facial regions and precisely distinguish between wrinkle and non-wrinkle areas to reduce false positives. Also, there may be benefits to including the type of wrinkle (e.g., static vs. dynamic wrinkle) to each wrinkle in the facial image because treatment strategies often differ by the type in clinics [28,8,7]. Despite majority voting, the subjectivity in wrinkle annotation remains a challenge. Moving forward, we plan to collaborate with dermatologists for wrinkle annotation and explore techniques such as soft labeling to improve the reliability and trustworthiness of ground truth wrinkles.

7 Conclusion

We propose a two-stage learning strategy for facial wrinkle segmentation that leverages transfer learning from facial texture feature extraction. Specifically, the

model is pretrained using automatically generated weak wrinkle labels (masked texture maps) to learn general facial features such as contours and skin texture. The model is then finetuned with a smaller set of manually labeled wrinkle data to enhance segmentation performance. This method demonstrates both qualitatively and quantitatively superior results, achieving state-of-the-art performance. Consequently, it significantly reduces the time and cost of manual wrinkle labeling, offering potential benefits in cosmetic dermatology. Additionally, the pretraining method’s architecture-independent nature suggests its broad applicability to various segmentation models, making it valuable not only in facial wrinkle segmentation but also in other areas requiring the segmentation of thin objects where manual labeling is costly. To support ongoing research and reproducibility, we have made the FFHQ-Wrinkle dataset—the first publicly available dataset of its kind—accessible to the research community. This dataset comprises 1,000 manually labeled wrinkle images and 50,000 weakly labeled images. By sharing this dataset, we aim to facilitate the development of more advanced wrinkle detection models and promote further advancements in this field.

Acknowledgements The authors appreciate Dr. Ik Jun Moon, a dermatologist at Asan Medical Center, for sharing invaluable insights and feedback from a dermatological perspective. This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (MSIT) (RS-2024-00455720 & RS-2024-00338048), the National Institute of Health(NIH) research project (2024ER040700), the National Supercomputing Center with supercomputing resources including technical support (KSC-2024-CRE-0021), Hankuk University of Foreign Studies Research Fund of 2024, the artificial intelligence semiconductor support program to nurture the best talents (IITP(2024)-RS-2023-00253914) grant funded by the Korea government, and the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024(RS-2024-00332210).

References

1. Allemann, I.B., Baumann, L.: Hyaluronic acid gel (juvéderm™) preparations in the treatment of facial wrinkles and folds. *Clinical interventions in aging* **3**(4), 629–634 (2008)
2. 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, Neuropsychology, and Cognition* **17**(4), 406–421 (2010)
3. Brempong, E.A., Kornblith, S., Chen, T., Parmar, N., Minderer, M., Norouzi, M.: Denoising pretraining for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4175–4186 (2022)
4. Chen, J., He, M., Cai, W.: Facial wrinkle detection with multiscale spatial feature fusion based on image enhancement and asff-seunet. *Electronics* **12**(24), 4897 (2023)

5. Crum, W.R., Camara, O., Hill, D.L.: Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging* **25**(11), 1451–1461 (2006)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
7. Gao, L., Song, W., Qian, L., Zhang, J., Li, K., Yang, J., Wang, G.: Clinical efficacy of different therapeutic modes of co2 fractional laser for treatment of static periorbital wrinkles in asian skin. *Journal of Cosmetic Dermatology* **21**(3), 1045–1050 (2022)
8. Goldman, A., et al.: Hyaluronic acid dermal fillers: Safety and efficacy for the treatment of wrinkles, aging skin, body sculpturing and medical conditions. *Clinical Medicine Reviews in Therapeutics* **3** (2011)
9. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *International MICCAI Brainlesion Workshop*. pp. 272–284. Springer (2021)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
11. Ji, Z., Shen, Y., Ma, C., Gao, M.: Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. pp. 175–183. Springer (2019)
12. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4401–4410 (2019)
13. Kim, K., Choi, Y.H., Hwang, E.: Wrinkle feature-based skin age estimation scheme. In: *2009 IEEE International Conference on Multimedia and Expo*. pp. 1222–1225. IEEE (2009)
14. 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)*. pp. 383–388. IEEE (2022)
15. Kim, S., Yoon, H., Lee, J., Yoo, S.: Facial wrinkle segmentation using weighted deep supervision and semi-automatic labeling. *Artificial Intelligence in Medicine* **145**, 102679 (2023)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 10012–10022 (2021)
17. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
19. Luu, K., Dai Bui, T., Suen, C.Y., Ricanek, K.: Combined local and holistic facial features for age-determination. In: *2010 11th International Conference on Control Automation Robotics & Vision*. pp. 900–904. IEEE (2010)
20. Ng, C.C., Yap, M.H., Cheng, Y.T., Hsu, G.S.: Hybrid ageing patterns for face age estimation. *Image and Vision Computing* **69**, 92–102 (2018)

21. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 ieee international conference on neural networks (ICNN'94). vol. 1, pp. 55–60. IEEE (1994)
22. Or-El, R., Sengupta, S., Fried, O., Shechtman, E., Kemelmacher-Shlizerman, I.: Lifespan age transformation synthesis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 739–755. Springer (2020)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
25. Sabina, U., Whangbo, T.K.: Edge-based effective active appearance model for real-time wrinkle detection. Skin Research and Technology **27**(3), 444–452 (2021)
26. Sabina, U., Whangbo, T.K.: Nasolabial wrinkle segmentation based on nested convolutional neural network. In: 2021 International Conference on Information and Communication Technology Convergence (ICTC). pp. 483–485. IEEE (2021)
27. Satriyasa, B.K.: Botulinum toxin (botox) a for reducing the appearance of facial wrinkles: a literature review of clinical use and pharmacological aspect. Clinical, cosmetic and investigational dermatology pp. 223–228 (2019)
28. Small, R.: Botulinum toxin injection for facial wrinkles. American family physician **90**(3), 168–175 (2014)
29. Warren, R., Gartstein, V., Kligman, A.M., Montagna, W., Allendorf, R.A., Ridder, G.M.: Age, sunlight, and facial skin: a histologic and quantitative study. Journal of the American Academy of Dermatology **25**(5), 751–760 (1991)
30. Wilder-Smith, E.P.: Stimulated skin wrinkling as an indicator of limb sympathetic function. Clinical Neurophysiology **126**(1), 10–16 (2015)
31. Wu, Y., Kalra, P., Thalmann, N.M.: Simulation of static and dynamic wrinkles of skin. In: Proceedings Computer Animation'96. pp. 90–97. IEEE (1996)
32. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
33. Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W.: Camel: A weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE/CVF International Conference on computer vision. pp. 10682–10691 (2019)
34. Yang, M.Y., Shen, Q.L., Xu, D.T., Sun, X.L., Wu, Q.B.: Striped wrinet: Automatic wrinkle segmentation based on striped attention module. Biomedical Signal Processing and Control **90**, 105817 (2024)
35. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018)
36. Zhou, Z.H.: A brief introduction to weakly supervised learning. National science review **5**(1), 44–53 (2018)