Article:
Liao, Jinpeng, Zhang, Tianyu, Li, Chunhui et al. (2024) LS-Net:lightweight segmentation network for dermatological epidermal segmentation in optical coherence tomography imaging. Biomedical Optics Express. pp. 5723-5738. ISSN: 2156-7085

https://doi.org/10.1364/BOE.529662

Research Article | Vol. 15, No. 10 / 1 Oct 2024 / Biomedical Optics Express 5723
Biomedical Optics EXPRESS

LS-Net: lightweight segmentation network for dermatological epidermal segmentation in optical coherence tomography imaging

JINPENG LIAO, TIANYU ZHANG, CHUNHUI LI,* AND ZHIHONG HUANG

University of Dundee, School of Science and Engineering, Dundee, United Kingdom
*c.li@dundee.ac.uk

Abstract: Optical coherence tomography (OCT) can be an important tool for non-invasive dermatological evaluation, providing useful data on epidermal integrity for diagnosing skin diseases. Despite its benefits, OCT's utility is limited by the challenges of accurate, fast epidermal segmentation due to the skin morphological diversity. To address this, we introduce a lightweight segmentation network (LS-Net), a novel deep learning model that combines the robust local feature extraction abilities of Convolution Neural Network and the long-term information processing capabilities of Vision Transformer. LS-Net has a depth-wise convolutional transformer for enhanced spatial contextualization and a squeeze-and-excitation block for feature recalibration, ensuring precise segmentation while maintaining computational efficiency. Our network outperforms existing methods, demonstrating high segmentation accuracy (mean Dice: 0.9624 and mean IoU: 0.9468) with significantly reduced computational demands (floating point operations: 1.131 G). We further validate LS-Net on our acquired dataset, showing its effectiveness in various skin sites (e.g., face, palm) under realistic clinical conditions. This model promises to enhance the diagnostic capabilities of OCT, making it a valuable tool for dermatological practice.

1. Introduction

Skin, the largest organ of the body, comprises three distinct layers: the epidermis, dermis, and subcutaneous tissue. It serves as the primary barrier against pathogens, ultraviolet radiation, and mechanical damage [1]. The integrity and functionality of the epidermis, mainly reflected in its thickness (ET), are crucial features for preventing skin fissures and ulcers. While the skin pathological conditions can influence the ET, the variations of ET are mainly decided by the different skin sites [2].

The golden standard for examining ET, e.g., biopsy, is invasive, non-repeatable, and often results in bleeding, scarring, and pain. In contrast, optical coherence tomography (OCT), a non-invasive, label-free imaging modality, provides a non-invasive, real-time, *in-vivo* assessment with axial resolutions around 10 µm and depth information up to 2 mm [3]. The efficacy of OCT in diagnosing skin conditions related to ET changes, such as skin cancer [4–6], skin acne [7], and inflammation [8,9], emphasizes its clinical value. However, manual annotations of OCT images require significant expertise and time [10], due to complex morphological changes in skin sites. Conventional methods for the OCT epidermis-dermis junction (EDJ), such as shapelet-based [11] and intensity-based [12] methods, heavily rely on image quality, struggling with image artifacts, and noise [13]. Moreover, compared to the easily distinguished sites like fingertips and palm, the forearm, neck, face, and wrist sites that have indistinct EDJ intensity/gray signals and morphological features are always difficult segmentation by the conventional methods. Clinical OCT imaging with handheld devices faces challenges like inconsistent angles, pressures, and

#529662 https://doi.org/10.1364/BOE.529662
Journal © 2024 Received 8 May 2024; revised 1 Jul 2024; accepted 31 Jul 2024; published 6 Sep 2024

Biomedical Optics EXPRESS

distances, affecting image quality [14]. Thus, it is essential to develop an efficient, automated, and accurate segmentation method for the rapid, real-time analysis of epidermal thickness.

Recent advancements in deep learning, particularly convolutional neural networks (CNNs) like U-Net [15], have improved the segmentation of the epidermal layer in OCT images [13,16–19]. Kepp et al. [20] further proposed a densely connected (DC)U-Net, which utilized densely connected blocks for enhanced feature reuse, demonstrating superior performance in mouse skin layer segmentation. However, this study focused on mouse skin, and for dermatological applications, it is crucial that models relearn the unique characteristics of human skin structure rather than solely focusing on image segmentation.

Despite these advances, CNN models struggle with capturing long-term dependencies due to their limited receptive fields and localized feature extraction mechanisms. In contrast, vision transformer (ViT) [21] provides a promising alternative with its ability to address long-term dependencies by self-attention mechanism, proving useful in OCT image processing tasks like image segmentation [22,23], reconstruction [24–27], and classification [28,29]. However, the self-attention mechanism also introduces significant computational complexity and necessitates large model sizes to achieve enhanced performance, making them less practical for medical imaging applications [30]. Moreover, distinct from CNNs, ViTs do not inherently account for the spatial relationships between pixels, thereby requiring larger datasets for effective training. This makes ViT less practical for medical imaging due to common resource constraints. Thus, there is a crucial need for developing smaller, more efficient models that do not compromise performance despite reduced computational resources and dataset availability.

To address these limitations, we propose a novel model, the lightweight segmentation (LS)-Net, which integrates the strengths of CNNs and ViTs. It includes a depth-wise convolutional transformer [31] to capture spatial relationships and a fusion layer incorporating a squeeze-and-excitation (SE) block [32] for enhanced feature integration. LS-Net is designed to be resource-efficient, containing only ∼0.5 M parameters and requiring ∼1.1 G floating-point operations, making it suitable for situations with limited computational resources.

Consequently, our study contributions are: (1) We propose LS-Net for efficient and accurate OCT-based segmentation of the skin epidermis, demonstrating superior performance compared to state-of-the-art networks. (2) We develop an intensity-based segmentation algorithm that generates substantial pseudo-data for pre-training, helping overcome training convergence and stability issues typical in ViT models. (3) Through ablation studies, we assess the effectiveness of the proposed LS-Net, focusing on network size, decoder heads, the feature fusion layer, and the strategy of using pseudo-data for pre-training. (4) We evaluate the ability of LS-Net to measure epidermal thickness, improving the practical utility of OCT in dermatology.

2. Related works

In the realm of medical image segmentation (MIS), the deep learning models can be divided into two main distinct architectures: CNNs and ViTs. During the early progress, FCN [33] stand as a cornerstone by introducing pixel-wise prediction capabilities in image segmentation. With a symmetrical architecture and long skip connection, U-Net has emerged as a widely adopted model due to its effective multi-scale feature fusion, setting a baseline in MIS.

By introducing ViT into U-Net, TransUNet [22] incorporates the self-attention mechanism of ViT, which has a global receptive field, enhancing the ability of long-term information processing compared to traditional CNNs. To further address efficiency concerns, Swin-UNet [23] uses shifted window mechanisms, which improve feature extraction scalability and computational efficiency across various scales, outperforming TransUNet in both aspects. Additionally, SegFormer [34] used an efficient self-attention that utilize convolution layer for query and key sequences reduction processing [35], decreasing self-attention complexity while preserving high performance in MIS. Despite these advancements, the compression techniques used to streamline

Biomedical Optics EXPRESS

self-attention in models could potentially compromise the retention of critical details, which is essential for precise segmentation of the epidermis-dermis junction.

3. Method

The architecture of our proposed Lightweight Segmentation Network (LS-Net) is depicted in Fig. 1. LS-Net simplifies the U-Net architecture into a compact encoder-decoder structure for multi-scale feature extraction. It has fewer parameters and less computational demand than U-Net, significantly reducing the network depth and the resources for inference and training. The encoder consists of depth-wise convolutional transformer (refer as DWCT in the following sections) blocks and overlapped patch extraction (OPE) layers. Although DWCT blocks offer efficient feature extraction, LS-Net lacks depth; we mitigate this through a novel fusion layer in the decoder, incorporating a SE-block for effective shallow-to-deep feature integration. The whole network is described in detail in the following sections.

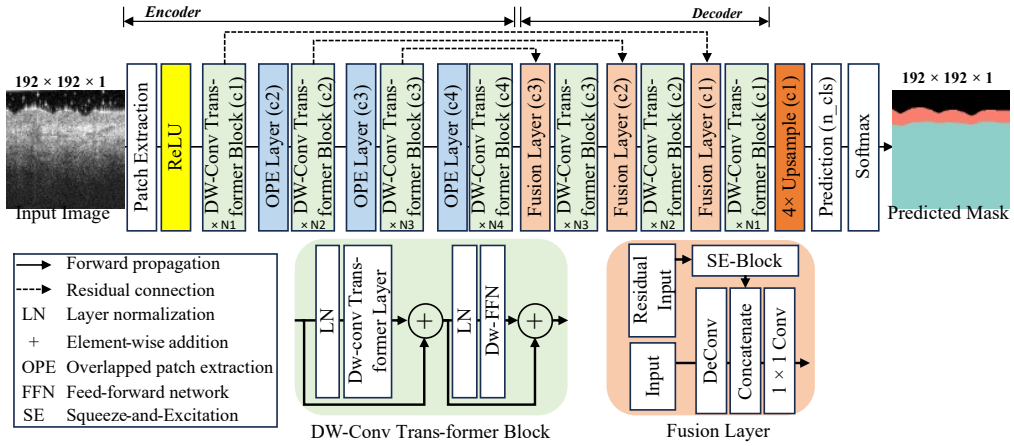

Fig. 1. The proposed LS-Net architecture. c1-c4 means the number of channel dimensions of the feature at each stage. N1-N4 means the number of depth-wise convolutional transformer blocks at each stage. n_cls is the number of classes. The prediction layer is set with a 1×1 convolution layer.

3.1. Overlapped patch extraction layer

Distinct from the fixed patch size and non-overlapping approach typical in ViT-like networks [21], we use overlapped patch extraction (OPE) layer to gradually obtain the patches at multi-scales in four stages. The OPE layer can offer the advantage of capturing local spatial relationships and inherent translation invariance, leading to richer feature representations that enhance model robustness and performance.

In LS-Net, the patch extraction layer and OPE layer are implemented with convolution layer with different setting of strides (S), kernel size (K), and output channel size (C). Assume the input of gray image has a shape of $\mathrm{H} \times \mathrm{W} \times 1$ (H: height; W: width), the first patch extraction layer ($K = 7$, $S = 4$, $C = 16$) split the image into patches of size 4×4, outputting image patches of dimensions $\mathrm{H}/4 \times \mathrm{W}/4 \times 16$. The setups of the following OPE layers were $K = 3$, $S = 2$, and $\mathrm{C}=\mathrm{c_i}$, where $i \in \{2, 3, 4\}$ and the implementation of c_i is in section 4.3. With OPE layers, the extracted feature was gradually downsampled with resolution $\mathrm{H}/2^{\mathrm{i}+1} \times \mathrm{W}/2^{\mathrm{i}+1} \times \mathrm{c_i}$, where $i \in \{2, 3, 4\}$.

Biomedical Optics EXPRESS

3.2. Depth-wise convolutional transformer (DWCT) block

As demonstrated in Fig. 2, the DWCT block integrates the DWCT (Fig. 2(A)) and depth-wise FFN (Fig. 2(C)), and Fig. 2(B) is the self-attention mechanism utilized in DWCT.

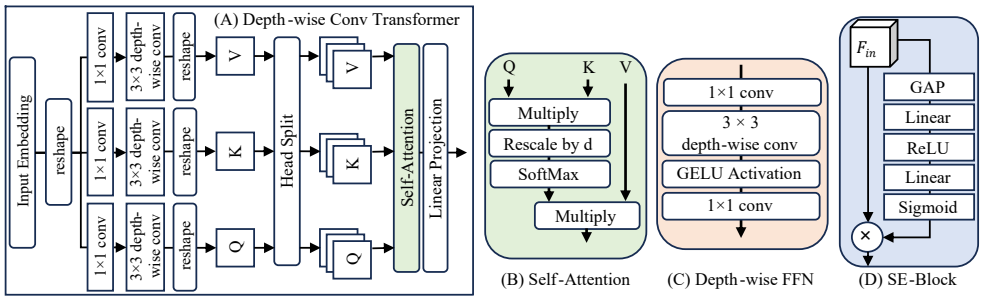

Fig. 2. The schematic of the component in the LS-Net. (A) The depth-wise convolutional transformer layer. (B) The self-attention mechanism. (C) Depth-wise feed forward network. (D) Squeeze-and-Excitation (SE)-block. Conv: convolution, GAP: global average pooling.

3.2.1. Depth-wise convolutional transformer

In the vanilla Transformer [21,22], the image patches are considered as 1D sequences, serving as the input of the self-attention (SA) mechanism input, and can be described as (1):

$$\mathrm{SA}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\mathrm{d}}\right)\mathrm{V} \tag{1}$$

where d is a numerical value of $\sqrt{dimension\ of\ Q}$, T is the transposing operation. Q (query), K (key), and V (value) are 1D sequences generated based on image patches with the linear projection layers. With SA, Transformer has an advantage over CNNs since it considers the information among all feature points, providing long-term information and a global receptive field. However, the vanilla Transformer does not consider the spatial relationship between the patches. To alleviate this problem, as shown in Fig. 2(A), the DWCT integrates depth-wise convolutions with self-attention, which maintains the global receptive field advantage from transformer while imparting spatial contextual information. This configuration also preserves computational efficiency without compromising the spatial information. Taking X with a shape of $\mathrm{H}\times\mathrm{W}\times\mathrm{C}$ as the input, the forward processing of the DWCT can be written as:

$$\mathrm{Sequence}_{\mathrm{Q,\,K,V}} = \mathrm{Reshape}(\mathrm{Dw}(\mathrm{Pw}(\mathrm{X}))) \tag{2}$$

$$\hat{\mathrm{X}} = \mathrm{LP}(\mathrm{SA}(\mathrm{Q},\mathrm{K},\mathrm{V})) \tag{3}$$

where Dw is a depth-wise convolution layer with a kernel size of 3×3, Pw is a 2D convolution layer with an implementation of kernel size of 1×1, stride of 1, and filter size of C. LP is a linear projection layer with a hidden size of C. After the reshape operation in (2), the shape of Q, K, and V sequences are $\mathrm{N}\times\mathrm{C}$, where N is $\mathrm{H}\times\mathrm{W}$. A head split operation is then applied to each sequence to reshape the sequence from $\mathrm{N}\times\mathrm{C}$ to $\mathrm{M}\times\mathrm{N/M}\times\mathrm{C}$, where M is the number of heads for multi-head self-attention (MHSA). After the MHSA processing, the output from the SA(Q, K, V) in (3) is a sequence with a shape of $\mathrm{M}\times\mathrm{N/M}\times\mathrm{C}$. The sequence is then reshaped to $\mathrm{N}\times\mathrm{C}$. Finally, after processing by LP layer, the output of (3), $\hat{\mathrm{X}}$, has a shape of $\mathrm{N}\times\mathrm{C}$.

Biomedical Optics EXPRESS

3.2.2. Depth-wise feed-forward network

Following the results of Xie et al. [34], that the convolution layer can be embedded in the feed-forward network (FFN) to provide better segmentation in the Transformer while maintaining the model efficiency, we introduce a depth-wise (DW)-FFN. DW-FFN (Fig. 2(C)) utilizes point-wise convolution and depth-wise convolution layers to replace the linear projection layers in the FFN, enhancing the ability to capture and process patterns, especially in understanding spatial relationships among local features. Taking the input is $\hat{X}$ from (3), the Dw-FFN is:

$$Y = Pw_2(GELU(Dw(Pw_1(\hat{X})))) \tag{4}$$

The shape of the input $\hat{X}$ is $N \times C$. Pw_1 and Pw_2 are convolution layers implemented with a kernel size of 1×1, stride of 1, while Pw_1 has a filter size of 4C, and Pw_2 has a filter size of C. Dw is a depth-wise convolution layer with a kernel size of 3×3.

3.3. Fusion layer

In the decoder, we proposed a fusion layer that consists of SE-block, deconvolution layer, and a 1×1 convolution layer. This assembly not only upsample feature maps but also better reuse the features from encoder section. As a plug-and-play module, the SE block enhances feature extraction by mapping channel relationships in convolutional features, thereby improving the representation of complex patterns. Furthermore, SE blocks improve network performance by adaptively recalibrating functions, adding little computational cost and requiring no major architectural changes. Taking X as the input, and $X_{residual}$ as the residual input from the encoder, the fusion layer can be written as:

$$out = Conv_{1\times1}(Concat(F_{SE}(X_{residual}),\ DeConv(X))) \tag{5}$$

where DeConv is a deconvolution layer with a kernel size of 3×3, stride of 2, and the filter size is the same as the input X, and F_{SE} is the SE-block mentioned in Fig. 2(D), $X_{residual}$ is the residual input from the previous encoder section. Concat is a feature concatenate layer, and $Conv_{1\times1}$ is a convolution layer with a kernel size of 1×1, stride of 1, filter size is implemented as section 4.3 mentioned (c_i, where $i \in \{1, 2, 3, 4\}$).

4. Experiments

4.1. Swept-source OCT system and data acquisition

A lab-built swept-source OCT (SSOCT) system was utilized to non-invasively acquire the data of skin structure with a hand-held probe (Fig. 3). More details of the system were described in [36]. The study was approved by the School of Science and Engineering Research Ethics Committee of the University of Dundee, which also conformed to the tenets of the Declaration of Helsinki. Before data collection, informed consent was obtained from all 36 participants, aged between 20 and 40, with no reported skin conditions. In the skin OCT data collection, a flexible hand-held scan probe was utilized to acquire the data from various skin sites including the palm, hand back, fingertip, wrist, face, neck, and forearm. To increase the amount of data, each location was collected thrice with minor positional adjustments. After the exclusion of scans with significant motion artifacts or bad quality, our dataset consists of 353 OCT data. (palm: 58; neck: 40; forearm: 83; fingertip: 48; face: 35; back of hand: 55; wrist: 34).

In terms of scanning protocol for data acquisition, one OCT scan can acquire data with a size of $6 \times 600 \times 600 \times 384$ (number of repetitions $\times$ X-transverse axis $\times$ Y-transverse axis $\times$ Z-axial axis). The spatial interval in the transverse axis is ~8.6 μm/pixel and theoretically ~7.4 μm/pixel (in air) in the axial axis. The field of view is 5.16 mm^2. We then applied the fast Fourier Transform (FFT) in OCT raw data pre-processing to convert spectral data into spatial information,

Biomedical Optics EXPRESS

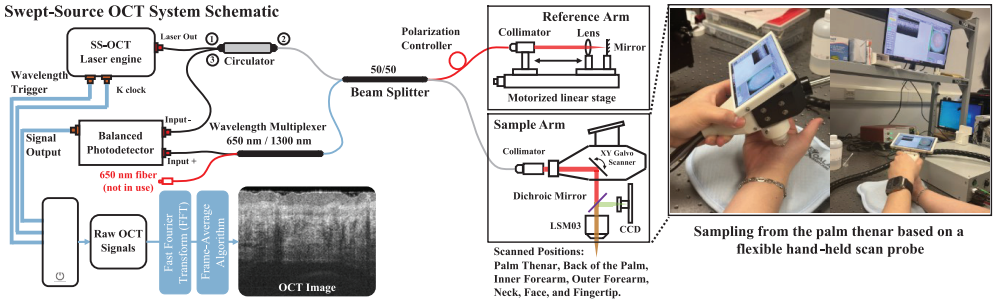

Fig. 3. The system schematic of the SSOCT system in this study. A demonstration of the hand-held scan probe for flexible data acquisition is shown on the right side of figure. The swept-source laser (SL132120 from Thorlabs Inc.) in our system has a wavelength of 1310 nm and a bandwidth of 100 nm. The A-scan swept rate is 200kHz.

thereby obtaining OCT volumes that contain high-resolution depth information. Since each OCT data set has six repetitions of OCT scans, we applied a frame-averaging algorithm [37] to reduce the speckle noise in the OCT volumes. Finally, we obtained 353 OCT volumes, each with dimensions of $600 \times 600 \times 384$ for further analysis.

4.2. Segmentation mask generation

For training, validation, and testing, we partitioned the 353 OCT volumes into two sets: 288 volumes from 25 participants were designated for the training and validation stages, and 65 volumes from 11 participants were reserved for testing. Our strategy consists of sampling every third frame from each volume to exploit wider structural diversity and avoid overfitting due to similar consecutive frames. We totally obtained 57,600 B-frames for the training and validation stage, with an additional 65 independent volumes for model performance testing.

Figure 4 illustrates the two-fold process for data preparation and model training. Due to the resource-intensive nature of the precise skin OCT image annotation, we adopted an intensity-based algorithm to automatically generate pseudo labels for pre-training stage (pseudocode can be found in Appendix in Algorithm 1). Although the accuracy of the pseudo label is not as high as the manual label, the pseudo label of OCT structure image is better for the model pre-train to learn the features of skin OCT structural images, rather than using the natural image. For data labelling, two experts annotated the OCT images from all OCT volumes and selected B-frames images every 30 frames in each OCT volume (1920 B-frames from 288 OCT volumes, and 430 B-frames from 65 OCT volumes in this study).

As shown in Fig. 4, a crop box with a shape of 192×192 is utilized to extract the image-label pairs for the model pre-train stage. In total, 172800 pairs of images and pseudo labels data are used to pre-train the model. Regarding the data used in fine-tuning stage, a series of crop boxes with a shape of 64^2, 80^2, 112^2, 144^2, and 192^2 are used to extract the image patches from the B-frames. Finally, 119040 pairs of images and manual labels data are used to fine-tuning the model.

4.3. Implementation details

The model used in this study were built and trained based on TensorFlow 2.9.0 [38]. We used the Adam optimizer [39] with an initial learning rate of 0.001 and a momentum of 0.9 for model optimization. The cross-entropy loss function was used to calculate the training loss between the provided mask and the model prediction. The number of training epochs was set as 400, and the batch size was set to 128. An NVIDIA RTX A6000 with 48 GB memory was used to facilitate the training of the model. To improve the model robustness, we applied random right and left

Biomedical Optics EXPRESS

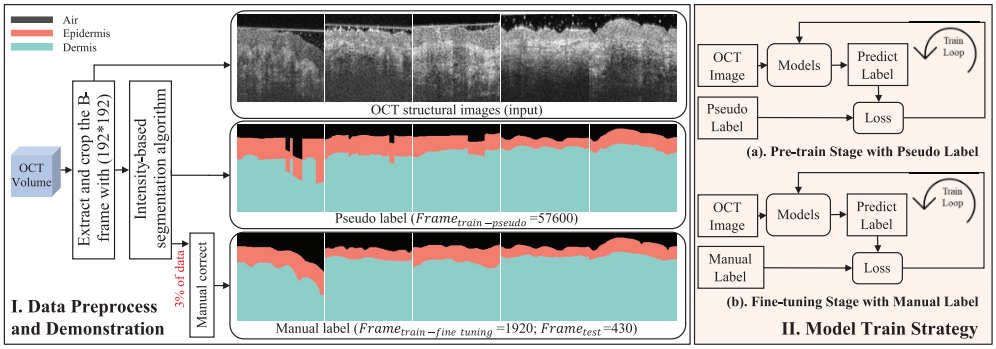

Fig. 4. (Stage-I) The pipeline of data pre-processing and mask generation. (Stage-II) The demonstration of the model pre-train stage (II-a) and fine-tuning stage (II-b).

image flipping as a data augmentation technique during the training. An early stopping strategy was used to prevent overfitting when the validation loss is not decreased over 40 consecutive epochs, and the model weights with the lowest loss were then saved.

Figure 4 stage-II illustrates the training strategies utilized in this study, including pre-train stage and fine-tuning stage. Apart from the data usage, these two training stages utilize the same training epochs and early stopping strategies as mentioned above.

Regarding the initialization of the proposed LS-Net, as shown in Fig. 1, the filter sizes (c_i, where $i \in \{1, 2, 3, 4\}$) of the DWCT block, OPE layer, and fusion layer at each stage are {16, 32, 64, 64}. In the setting of DWCT block, the number of head (H_i, where $i \in \{1, 2, 3, 4\}$) at each stage are {1, 2, 4, 4}, and the number of DWCT block (N_i, where $i \in \{1, 2, 3, 4\}$) at each stage are {1, 1, 2, 2}. The 4x upsample layer is a deconvolution layer with a kernel size of 7×7, a stride of 4, and a filter size of c_1. In terms of the prediction layer, the kernel size is 1×1, the stride is 1, and the filter size is 3, which is equal to the number of classes.

4.4. Evaluation metrics

To quantify the segmentation accuracy of various methods, we utilized six metrics in this study, including accuracy (Acc), specificity (Spe), sensitivity (Sen), precision (Pre), the mean dice similarity coefficient (mDice), and the mean intersection over union (mIoU). Those metrics can be formulated as:

$$\mathrm{mIoU} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\mathrm{TP_i}}{\mathrm{TP_i} + \mathrm{FP_i} + \mathrm{FN_i}} \tag{6}$$

$$\mathrm{mDice} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i} \tag{7}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{9}$$

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{11}$$

Biomedical Optics EXPRESS

where TP (/TN) is the true position (/negative) and represents the number of pixels correctly predicted and labeled as positive (/negative). Conversely, the number of pixels that are incorrectly given a positive (/negative) label is called FP(/FT). N is the number of classes.

4.5. Comparison to state-of-the-art methods

Given the novelty of our dataset, direct comparisons with the existing methods are not feasible. Therefore, we evaluate the performance of our LS-Net for skin layer segmentation with various models that are proposed for related studies (e.g., medical image segmentation). Those models include high-efficiency and high-performance medical image segmentation models (T-Net [40], TransUNet [22], SwinUNet [23], Wave-Net [41], SegFormer (mit-b0 setup) [34], SHFormer [42], and CENet [43]), and methods for OCT-based skin layer segmentation (DCU-Net [20], UNet [15,17,44]). The training details and training strategies of the compare-used models are the same as the proposed implementation details mentioned in section 4.3, to reduce the influence from the training details and hardware. The results are shown in Table 1.

Table 1. Quantitative Comparison (mean ± standard deviation) with State-of-the-art Methods.

Method	mDice↑	mIoU↑	Sen↑	Spe↑	Acc↑	Pre↑	FLOPs	Params
UNet	0.9573 ± 0.015	0.9394 ± 0.027	0.9866 ± 0.006	0.9933 ± 0.003	0.9866 ± 0.006	0.9866 ± 0.006	60.00 G	34.56 M
DCU-Net	0.9553 ± 0.013	0.9393 ± 0.023	0.9863 ± 0.006	0.9932 ± 0.003	0.9863 ± 0.006	0.9863 ± 0.006	13.14 G	6.106 M
TransUNet	0.9499 ± 0.018	0.9294 ± 0.030	0.9828 ± 0.008	0.9914 ± 0.004	0.9828 ± 0.008	0.9828 ± 0.008	23.01 G	52.35 M
SwinUNet	0.9465 ± 0.021	0.9233 ± 0.036	0.9822 ± 0.011	0.9911 ± 0.005	0.9822 ± 0.011	0.9822 ± 0.011	16.30 G	50.28 M
T-Net	0.9506 ± 0.033	0.9295 ± 0.048	0.9830 ± 0.015	0.9915 ± 0.008	0.9829 ± 0.015	0.9830 ± 0.015	0.478 G	0.045 M
CENet	0.9617 ± 0.015	0.9418 ± 0.026	0.9869 ± 0.006	0.9934 ± 0.003	0.9869 ± 0.006	0.9869 ± 0.006	10.89 G	15.62 M
Wave-Net	0.9493 ± 0.020	0.9273 ± 0.033	0.9809 ± 0.010	0.9904 ± 0.005	0.9809 ± 0.010	0.9809 ± 0.010	150.7 G	7.882 M
SegFormer	0.9593 ± 0.015	0.9404 ± 0.025	0.9865 ± 0.006	0.9932 ± 0.003	0.9865 ± 0.006	0.9865 ± 0.006	1.932 G	3.702 M
SHFormer	0.9594 ± 0.015	0.9432 ± 0.023	0.9872 ± 0.005	0.9936 ± 0.003	0.9872 ± 0.005	0.9872 ± 0.005	0.597 G	2.817 M
LS-Net	**0.9624 ± 0.014**	**0.9468 ± 0.023**	**0.9882 ± 0.005**	**0.9941 ± 0.003**	**0.9882 ± 0.005**	**0.9882 ± 0.005**	1.131 G	0.507 M

The results reveal that our proposed LS-Net has demonstrated exceptional performance across multiple metrics, including mDice (0.9624) and mIoU (0.9468), outperforming other methods. Notably, it also exhibited a high sensitivity (0.9882) and specificity (0.9941), which indicates robustness in identifying true positives and negatives. The accuracy further solidified its reliability with a score of 0.9882. Moreover, LS-Net maintained computational efficiency, with a reduced number of floating points operation (FLOPs) at 1.131 G, which is significantly lower than several other models, and a modest number of parameters totaling 0.507 M.

The visual comparisons from Fig. 5 to Fig. 7 present the segmentation performance of various deep learning models on cross-sectional images of skin structures from the face, palm, and forearm, respectively. All figures are scalable vector graphic format, and please feel free to zoom in for detailed comparison. These anatomical regions exhibit diverse textural and contrast characteristics, serving as a challenging test for accurate segmentation. Among the models evaluated, LS-Net consistently exhibits superior performance in matching the expert-annotated ground truth across all three anatomical locations.

In Fig. 5 (face), LS-Net achieves the highest Intersection over Union (IoU) of 0.914 and Dice coefficient of 0.938, indicating a remarkable overlap with the manual labels. Similarly, for the palm structure in Fig. 6, LS-Net closely approximates the expert annotations, attaining the best IoU of 0.947 and Dice of 0.966. Even in the challenging case of the forearm (Fig. 7), where contrast variations and image noise can negatively influence segmentation, LS-Net secures an impressive IoU of 0.918 and Dice of 0.945, outperforming its counterparts.

Biomedical Optics EXPRESS

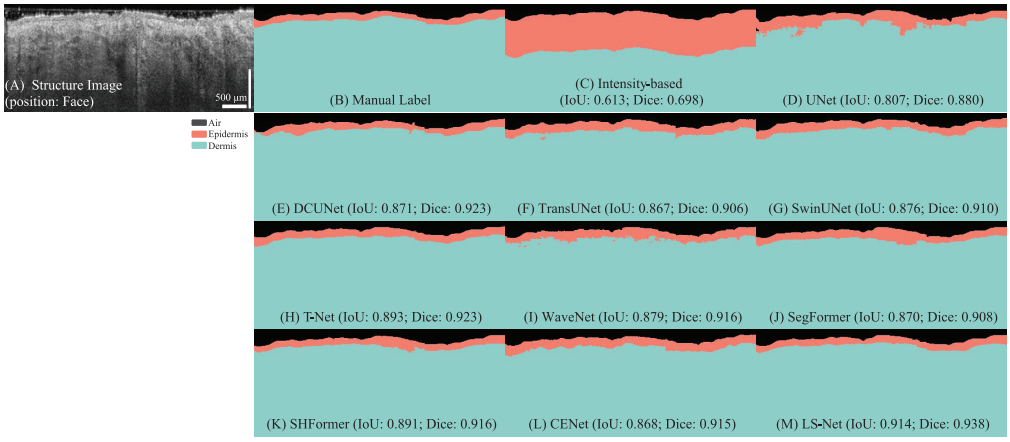

Fig. 5. Segmentation results from various methods (Position: Face). (A) Cross sectional structure image. (B) Expert annotations. (C) Pseudo label generated by intensity segmentation method. (D)-(M) are segmentation masks generated by various deep-learning models.

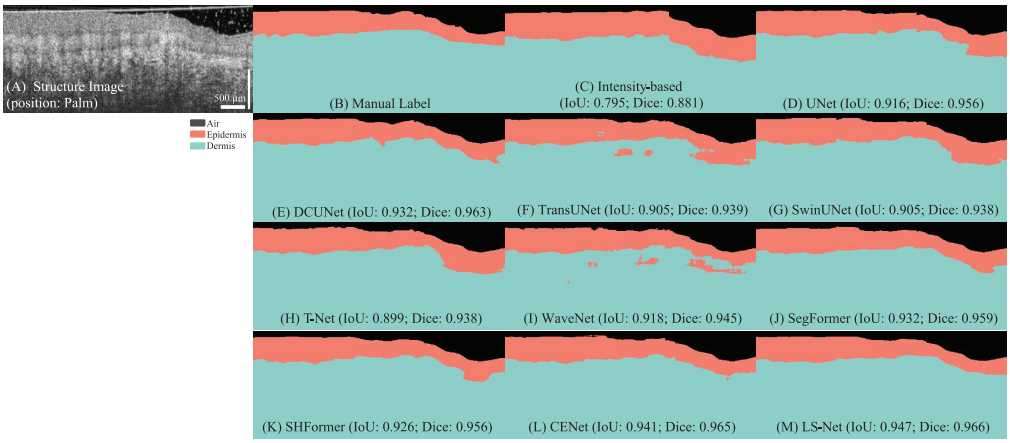

Fig. 6. Segmentation results from various methods (Position: Palm). (A) Cross sectional structure image. (B) Expert annotations. (C) Pseudo label generated by intensity segmentation method. (D)-(M) are segmentation masks generated by various deep-learning models.

4.6. Ablation Studies

4.6.1. Influence of network size

In this section, we evaluated the influence of the network size by implementing various filter sizes (c_i, where $i \in \{1, 2, 3, 4\}$) and attention head number (H_i, where $i \in \{1, 2, 3, 4\}$). Following the implementation details in section 4.3, we define the original setup is LS-Net-B. In terms of the LS-Net-S and LS-Net-L, the implementation for DWCT Blocks is shown in Table 2.

The results in Table 2 indicate that as the network scale increases, the performance metrics show a clear upward trend (mDice from 0.9553 to 0.9629, and mIoU from 0.9371 to 0.9444). However, this improved performance comes with increased model complexity, as evidenced by higher FLOPs (from 0.549 G to 2.686 G) and network parameters. Compared to the larger LS-Net-L model, the LS-Net-B configuration uses intermediate filter sizes and attention heads,

Biomedical Optics EXPRESS

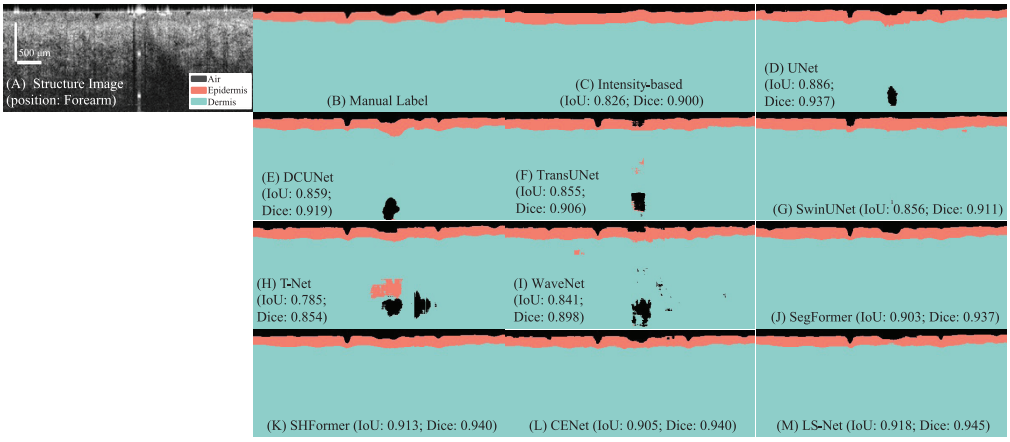

Fig. 7. Segmentation results from various methods (Position: Forearm). (A) Cross sectional structure image. (B) Expert annotations. (C) Pseudo label generated by intensity segmentation method. (D)-(M) are segmentation masks generated by various deep-learning models.

Table 2. Ablation study on the architecture

Model	Filter Size (c)	Heads (H)	mDice↑	mIoU↑	FLOPs	Params
LS-Net-S	8, 16, 32, 32	1, 2, 4, 4	0.9553 ± 0.020	0.9371 ± 0.032	0.549 G	0.1359 M
LS-Net-B	16, 32, 64, 64	1, 2, 4, 4	0.9624 ± 0.014	0.9468 ± 0.023	1.131 G	0.5069 M
LS-Net-L	32, 64, 128, 128	2, 4, 8, 8	0.9629 ± 0.014	0.9474 ± 0.024	2.686 G	1.9540 M

resulting in a slight performance decrease (mDice: $0.9624 < 0.9629$; mIoU: $0.9468 < 0.9474$), but a significant reduction in model complexity (FLOPs: $1.131\,\text{G} < 2.686\,\text{G}$). Compared to LS-Net-B, the smaller LS-Net-S model has lower computational requirements in terms of FLOPs ($0.549\,\text{G} < 1.131\,\text{G}$) and parameters ($0.1359\,\text{M} < 0.5069\,\text{M}$), but also has lower performance (mDice: $0.9553 < 0.9624$; mIoU: $0.9371 < 0.9468$). Despite the performance drop, LS-Net-S is potentially suitable for applications with limited resources.

4.6.2. Influence of SE-block

In section 3.3, we proposed a fusion layer with SE-Block to enhance the feature fusion between the shallow and deeper features for better segmentation performance. To quantify the effect of the SE-Block, in this experiment, we compare the performance of the fusion layer between the with (w) SE-Block and without (w/o) SE-Block.

The results in Table 3 shows that adding the SE-Block to the fusion layer can improve the segmentation performance (mDice: $0.9624 > 0.9595$; mIoU: $0.9468 > 0.9413$). Moreover, the inclusion of SE-Block also slightly improves the specificity and accuracy of model, reaching 0.9941 and 0.9882, respectively. These performance gains are accomplished with a negligible increase in computational complexity, as the FLOPs only marginally increase from 1.13136 G to 1.13149 G, and the number of parameters remains relatively stable (0.5054 M vs. 0.5069 M).

Table 3. Ablation study on the architecture of the design of the feature fusion layer

SE-Block	mDice↑	mIoU↑	Spe↑	Acc↑	FLOPs (G)	Params
w/o	0.9595 ± 0.015	0.9413 ± 0.026	0.9934 ± 0.003	0.9868 ± 0.006	1.13136	0.5054 M
w	0.9624 ± 0.014	0.9468 ± 0.023	0.9941 ± 0.003	0.9882 ± 0.005	1.13149	0.5069 M

Biomedical Optics EXPRESS

4.6.3. Influence of training strategies

In section 4.2 and 4.3, we proposed an intensity-based segmentation method to create a large amount of relatively low-accuracy pseudo segmentation data for model pre-training. To assess the impact of this training strategy, we compared the performance of LS-Net models trained with the proposed strategy versus those trained only on high-accuracy manual labels.

As shown in Table 4, the LS-Net models pre-trained with pseudo data exhibited superior performance across several evaluation metrics. The mean Dice score increased slightly from 0.9591 to 0.9624 and the mean IoU increased from 0.9425 to 0.9468, suggesting a more precise overlap between model predictions and ground truth. These results indicate that the LS-Net models pre-trained with pseudo segmentation data not only retained high levels of specificity and sensitivity, but also demonstrated improvements in overall accuracy, precision, and segmentation metrics compared to training with manual labels alone.

Table 4. Ablation study on the various training strategies

Pseudo Data	mDice↑	mIoU↑	Sen↑	Spe↑	Acc↑	Pre↑
√	0.9624 ± 0.014	0.9468 ± 0.023	0.9882 ± 0.005	0.9941 ± 0.003	0.9882 ± 0.005	0.9882 ± 0.005
×	0.9591 ± 0.015	0.9425 ± 0.026	0.9871 ± 0.007	0.9935 ± 0.003	0.9871 ± 0.007	0.9871 ± 0.007

4.6.4. Influence of different decoders

Recently, the MLP decoder has been widely used in medical image segmentation models to reduce network complexity and parameters [34,42]. However, the performance of these lightweight decoders has not been extensively evaluated on OCT-based skin layer segmentation tasks. To address this gap, we conducted an ablation study to evaluate the influence of various decoders when coupled with our LS-Net architecture. Hence, we selected three representative lightweight-design decoders (U-Net [15], SegFormer [34], and SHFormer [42]) and compared their performance against our proposed LS-Net decoder. As shown in Table 5, the LS-Net decoder outperformed the other decoders in both mDice (0.9624) and mIoU (0.9468) metrics.

Table 5. Ablation study on the various decoders

Decoder	mDice↑	mIoU↑	Spe↑	Acc↑	FLOPs	Params
LS-Net (**ours**)	**0.9624 ± 0.014**	**0.9468 ± 0.023**	**0.9941 ± 0.003**	**0.9882 ± 0.005**	1.1314 G	0.5069 M
SegFormer	0.9580 ± 0.014	0.9416 ± 0.022	0.9934 ± 0.003	0.9869 ± 0.006	0.5729 G	0.3541 M
U-Net	0.9574 ± 0.015	0.9394 ± 0.025	0.9932 ± 0.003	0.9863 ± 0.006	0.7225 G	0.5155 M
SHFormer	0.9594 ± 0.015	0.9420 ± 0.025	0.9936 ± 0.003	0.9872 ± 0.005	0.5284 G	0.3044 M

While the LS-Net decoder exhibited the highest FLOPs (1.1314 G) and parameter count (0.5069 M) among the decoders evaluated, this increased complexity is necessary to achieve superior segmentation performance. For instance, the LS-Net decoder outperformed U-Net by 0.005 in mDice and 0.0074 in mIoU, despite having higher computational requirements. Similarly, the LS-Net decoder surpassed SegFormer and SHFormer by 0.0044 and 0.003 in mDice, and 0.0052 and 0.0048 in mIoU, respectively. These performance gains demonstrate that the increased complexity of the LS-Net decoder is necessary to capture the intricate details of the skin layers and achieve more accurate segmentation.

4.7. Model inference complexity comparison

Figure 8 presents the comparison of the model inference complexity of various deep-learning models and their components under different batch size settings on GPU.

Biomedical Optics EXPRESS

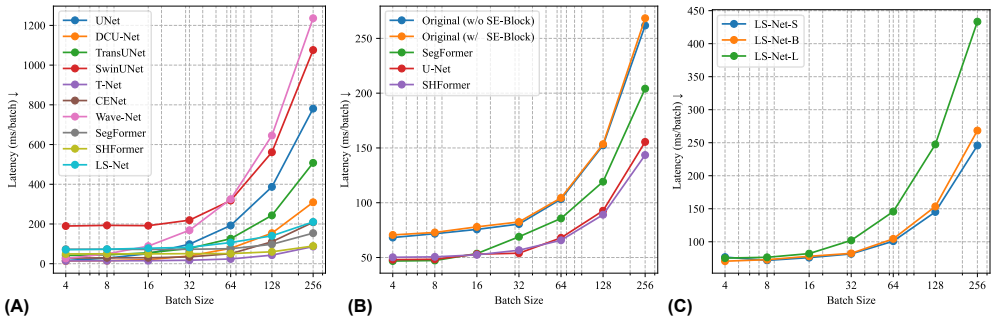

Fig. 8. Comparison of Model Inference Efficiency under Different Batch Size Settings on GPU. (A) The comparison in different models (B) The comparisons in the various decoder heads utilized in LS-Net. (C) Comparisons of the various sizes of the LS-Net.

Figure 8(A) is a comparison based on latency (in milliseconds) across various models. The graph demonstrates that as the batch size increases, the inference latency generally increases for all models, but the rate of increase varies among them. Among them, UNet, Wave-Net, and Swin-UNet show a steep increase in latency as the batch size grows, indicating a high computational complexity or less efficient batching. The increasing rate of latency in LS-Net, SegFormer, SHFormer, and T-Net is relatively slower, indicating better handling of larger batch sizes compared to U-Net.

Figure 8(B) is the comparison of the various decoder heads utilized in LS-Net. Similar to the FLOPs and parameters results shown in Table 5, the original LS-Net decoder has the highest latency and performance worst in computational complexity. Among them, the LS-Net with SHFormer decoder has the lowest latency, and the U-Net decoder has a similar performance.

Figure 8(C) is the comparison in terms of the size of LS-Net. The results show a direct relationship between the size of the LS-Net and latency, with larger models resulting in higher latency. Among them, LS-Net-S is the most efficient across all batch sizes. LS-Net-B has a similar latency with the LS-Net-S, while has better segmentation performance (in Table 2).

4.8. Epidermis thickness evaluation

The precision of epidermal thickness measurements is critical for dermatological diagnostics. In this study, the pixel size of the utilized SSOCT system is 8.74 μm/pixel in air. Considering the refractive index transition from air to skin tissue ranges between 1.36 and 1.44 [45], where

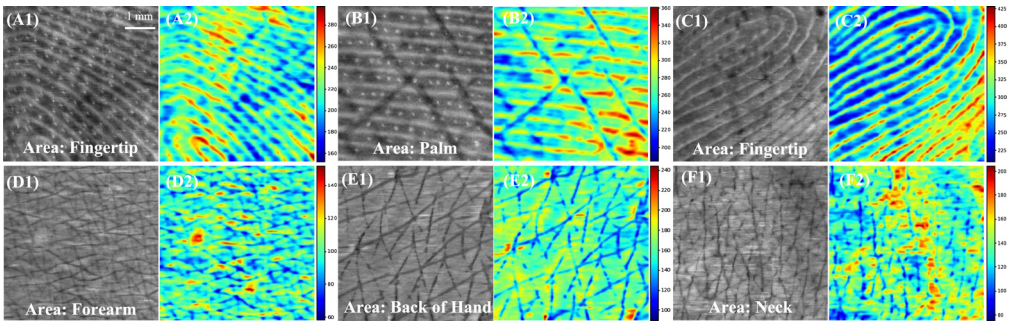

Fig. 9. A1 to F1 are *en face* images of skin epidermis layers with maximum intensity projection algorithm. A2 to F2 are epidermal thickness heatmaps with color bars. (The white scale bar applies to all images. The unit of the heatmap color bar is micro-meter.)

Biomedical Optics EXPRESS

we adopt a median value of 1.40, the pixel size of 6.24 μm/pixel in the skin tissue was used when calculating the epidermal thickness. Figure 9 shows the *en face* images (A1-F1) of the skin epidermis layer and the corresponding heatmaps of epidermal thickness (A2-F2) across different skin locations. The color in these heatmaps represents the variation of epidermal thicknesses. In Fig. 9 A2-F2, the proposed LS-Net can capture subtle physiological differences, such as fingerprints.

5. Discussion

In this study, we proposed a LS-Net, an innovative approach to OCT-based epidermal layer segmentation that effectively integrates the advantages of CNN and ViT. The LS-Net has the highest segmentation performance (mean Dice: 0.9468; mean IoU: 0.9624) with moderate computational demands (FLOPs:1.131 G) compared to existing state-of-the-art methods. One of the main contributions of LS-Net is the depth-wise convolutional transformer block, which enhances the ability to capture spatial relationship information while maintaining the global receptive field provided by self-attention mechanisms. This is essential in medical imaging because the precision of segmentation can directly influence diagnosis. Moreover, the SE-block in the fusion layers enables the recalibration of feature channels, thereby improving the representational capacity of the network without significantly increasing the parameters and model complexity. Besides, we proposed an intensity-based segmentation algorithm that can generate a large amount of pseudo data for model pre-trained, and the ablation results show that the pre-trained with pseudo data can improve the segmentation performance of the model.

As evident from the qualitative and quantitative results shown in Fig. 5 to Fig. 7, LS-Net demonstrates strong capability in handling the diverse textural and structural complexities inherent in skin OCT images. This has significant potential in clinical applications such as early detection and monitoring of skin conditions, including cancerous lesions and inflammatory diseases. The segmentation mask from LS-Net not only maintains structural integrity but also minimizes false positives and negatives. Notably, LS-Net performs well in segmenting forearm images (Fig. 7), since it can expertly handle the artifacts and high reflection challenges while preserving fine details of segmentation. The model reliably segments fine skin details, aiding early detection of conditions and treatment monitoring when integrated into diagnostic workflows.

However, the study has limitations. Firstly, the OCT imaging data collected in this study is from one OCT device in our lab, hence, the reliance on OCT imaging data may restrict the use of LS-Net in environments where such resources are scarce. Thus, the operational environments for this technology might be limited by the availability of OCT devices. Secondly, LS-Net is designed for computational efficiency, but it does face trade-offs between performance and computational demand when compared to models such as T-Net and SHFormer. Our ablation studies suggest that increasing model complexity to improve performance also increases computational demand and inference latency. This becomes particularly challenging in batch processing scenarios where swift processing is crucial, potentially limiting LS-Net's practicality in resource-limited settings that demand rapid and efficient processing. Thirdly, the study validated LS-Net on a dataset limited to individuals aged 20-40 with no reported skin conditions, which does not adequately reflect the diversity of clinical realities. It may not provide a comprehensive model's performance across different skin types, age groups, and pathologies. Therefore, additional validation studies incorporating a more diverse patient population are essential to thoroughly assess the model's effectiveness in varied clinical settings and to ensure its broader applicability.

Future research can focus on several areas to enhance the applicability and efficiency of LS-Net. First, the diversity of the training dataset should be broadened to include a more extensive range of skin types, ages, and pathological conditions. This expansion is crucial for enhancing the model's robustness and ensuring its generalizability across varied clinical environments. Second, exploring the integration of additional lightweight transformer models may further

Biomedical Optics EXPRESS

reduce computational requirements while maintaining high accuracy. Finally, real-world clinical trials will be critical to validate the practical benefits of LS-Net in healthcare settings, possibly incorporating feedback from dermatologists to further refine the model.

6. Conclusion

In conclusion, LS-Net represents a significant advancement in medical image segmentation, providing an efficient and accurate tool for non-invasive skin analysis that could facilitate fast and reliable clinical outcomes through enhanced diagnostic capabilities. By efficiently integrating CNNs and ViTs, LS-Net achieves high segmentation accuracy with low computational demands. The model has the best performance over the state-of-the-art methods, particularly in maintaining the accuracy of epidermal thickness measurements, which are crucial for diagnosing and monitoring various skin conditions.

Appendix: pseudocode of the intensity-based segmentation algorithm

Algorithm 1. Function: Intensity-based Segmentation Algorithm

Input: OCT cross-section image (bframe), Gaussian kernel size (ks_gaussian, (default: 7)), gamma values (γ1 for AEJ (default: 1.2), γ2 for EDJ (default:7))

1. Image Denoising:
- Threshold 'bframe' by setting pixels with a value less than the mean of 'bframe' to 0, to denoise the image.
2. AEJ Segmentation: # AEJ is air-epidermis-junction.
- Apply a Gaussian filter to 'bframe' with a kernel size of 'ks_gaussian'.
- Adjust the gamma of the filtered image using 'γ1'.
- Detect the AEJ line by finding the peak gradient position in each column after smoothing, indicating the transition in each column of the image.
3. EDJ Segmentation: # EDJ is epidermis-dermis-junction.
- Starting below the AEJ, enhance the contrast of the remaining image using 'γ2' to prepare for EDJ detection.
- Apply a Gaussian filter to the contrast-enhanced image with a kernel size of 'ks_gaussian'.
- Detect the EDJ line by locating the peak gradients in the adjusted region, similar to the AEJ detection but focused on the deeper part of the skin.

Output: AEJ line, EDJ line

Acknowledgments. Jinpeng would like to provide his greatest thanks to Prof. Zhihong Huang and Dr. Chunhui Li for their support.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data and Python code underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. F. F. Sahle, T. Gebre-Mariam, B. Dobner, *et al.*, "Skin diseases associated with the depletion of stratum corneum lipids and stratum corneum lipid substitution therapy," Skin Pharmacol. Physiol. **28**(1), 42–55 (2015).
2. D. A. Lintzeri, N. Karimian, U. Blume-Peytavi, *et al.*, "Epidermal thickness in healthy humans: a systematic review and meta-analysis," Acad. Dermatol. Venereol. **36**(8), 1191–1200 (2022).
3. W. Fujimoto James and G. Drexler, "Introduction to OCT," in *Optical Coherence Tomography: Technology and Applications*, J. G. Drexler, ed. (Springer International Publishing, 2015), pp. 3–64.
4. M. Mogensen, L. Thrane, T. M. Jørgensen, *et al.*, "Optical coherence tomography for imaging of skin and skin diseases," in *Seminars in Cutaneous Medicine and Surgery* (WB Saunders, 2009), 28(3), pp. 196–202.
5. A. B. E. Attia, S. Y. Chuah, D. Razansky, *et al.*, "Noninvasive real-time characterization of non-melanoma skin cancers with handheld optoacoustic probes," Photoacoustics **7**, 20–26 (2017).
6. A. Rajabi-Estarabadi, J. M. Bittar, C. Zheng, *et al.*, "Optical coherence tomography imaging of melanoma skin cancer," Lasers Med. Sci. **34**(2), 411–420 (2019).
7. U. Baran, Y. Li, W. J. Choi, *et al.*, "High resolution imaging of acne lesion development and scarring in human facial skin using OCT-based microangiography," Lasers Surg. Med. **47**(3), 231–238 (2015).

Biomedical Optics EXPRESS

8. A. J. Deegan, F. Talebi-Liasi, S. Song, *et al.*, "Optical coherence tomography angiography of normal skin and inflammatory dermatologic conditions," Lasers Surg. Med. **50**(3), 183–193 (2018).
9. Y.-J. Wang, J.-Y. Wang, and Y.-H. Wu, "Application of cellular resolution full-field optical coherence tomography in vivo for the diagnosis of skin tumours and inflammatory skin diseases: a pilot study," Dermatology **238**(1), 121–131 (2022).
10. S. Ud-Din, P. Foden, K. Stocking, *et al.*, "Objective assessment of dermal fibrosis in cutaneous scarring, using optical coherence tomography, high-frequency ultrasound and immunohistomorphometry of human skin," Br. J. Dermatol. **181**(4), 722–732 (2019).
11. J. Weissman, T. Hancewicz, and P. Kaplan, "Optical coherence tomography of skin for measurement of epidermal thickness by shapelet-based image analysis," Opt. Express **12**(23), 5760–5769 (2004).
12. Y. Hori, Y. Yasuno, S. Sakai, *et al.*, "Automatic characterization and segmentation of human skin using three-dimensional optical coherence tomography," Opt. Express **14**(5), 1862–1877 (2006).
13. Y. Ji, S. Yang, K. Zhou, *et al.*, "Deep-learning approach for automated thickness measurement of epithelial tissue and scab using optical coherence tomography," J. Biomed. Opt. **27**(01), 015002 (2022).
14. J. Liao, T. Zhang, Y. Zhang, *et al.*, "VET: vasculature extraction transformer for single-scan optical coherence tomography angiography," IEEE Trans. Biomed. Eng. **71**(4), 1179–1190 (2024).
15. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2015), pp. 234–241.
16. X. Liu, S. Ouellette, M. Jamgochian, *et al.*, "One-class machine learning classification of skin tissue based on manually scanned optical coherence tomography imaging," Sci. Rep. **13**(1), 867 (2023).
17. X. Liu, N. Chuchvara, Y. Liu, *et al.*, "Real-time deep learning assisted skin layer delineation in dermal optical coherence tomography," OSA Continuum **4**(7), 2008–2023 (2021).
18. R. Del Amor, S. Morales, A. Colomer, *et al.*, "Automatic segmentation of epidermis and hair follicles in optical coherence tomography images of normal skin by convolutional neural networks," Front. Med. **7**, 220 (2020).
19. Y. Lin, D. Li, W. Liu, *et al.*, "A measurement of epidermal thickness of fingertip skin from OCT images using convolutional neural network," J. Innov. Opt. Health Sci. **14**(01), 2140005 (2021).
20. T. Kepp, C. Droigk, M. Casper, *et al.*, "Segmentation of mouse skin layers in optical coherence tomography image data using deep convolutional neural networks," Biomed. Opt. Express **10**(7), 3484–3496 (2019).
21. A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations* (2020).
22. J. Chen, Y. Lu, Q. Yu, *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv, (2021).
23. H. Cao, Y. Wang, J. Chen, *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European Conference on Computer Vision* (Springer, 2022), pp. 205–218.
24. J. Liao, T. Zhang, C. Li, *et al.*, "U-shaped fusion convolutional transformer based workflow for fast optical coherence tomography angiography generation in lips," Biomed. Opt. Express **14**(11), 5583–5601 (2023).
25. J. Liao, S. Yang, T. Zhang, *et al.*, "Fast optical coherence tomography angiography image acquisition and reconstruction pipeline for skin application," Biomed. Opt. Express **14**(8), 3899–3913 (2023).
26. Z. Jiang, Z. Huang, B. Qiu, *et al.*, "Weakly supervised deep learning-based optical coherence tomography angiography," IEEE Trans. Med. Imaging **40**(2), 688–698 (2021).
27. X. Liu, Z. Huang, Z. Wang, *et al.*, "A deep learning based pipeline for optical coherence tomography angiography," J. Biophotonics **12**(10), e201900008 (2019).
28. S. Chen, Z. Wu, M. Li, *et al.*, "Fit-net: Feature interaction transformer network for pathologic myopia diagnosis," IEEE Trans. Med. Imaging **42**(9), 2524–2538 (2023).
29. J. He, J. Wang, Z. Han, *et al.*, "An interpretable transformer network for the retinal disease classification using optical coherence tomography," Sci. Rep. **13**(1), 3637 (2023).
30. H. Ren, H. Dai, Z. Dai, *et al.*, "Combiner: Full attention transformer with sparse computation cost," Adv. Neural. Inf. Process Syst. **34**, 22470–22482 (2021).
31. S. W. Zamir, A. Arora, S. Khan, *et al.*, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5728–5739.
32. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141.
33. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3431–3440.
34. E. Xie, W. Wang, Z. Yu, *et al.*, "SegFormer: Simple and efficient design for semantic segmentation with transformers," Adv. Neural. Inf. Process Syst. **34**, 12077–12090 (2021).
35. W. Wang, E. Xie, X. Li, *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 568–578.
36. J. Liao, C. Li, and Z. Huang, "A Lightweight Swin Transformer-Based Pipeline for Optical Coherence Tomography Image Denoising in Skin Application," Photonics **10**(4), 468 (2023).
37. W. Wu, O. Tan, R. R. Pappuru, *et al.*, "Assessment of frame-averaging algorithms in OCT image analysis," Ophthalmic Surg Lasers Imaging Retina **44**(2), 168–175 (2013).

Biomedical Optics EXPRESS

38. M. Abadi, P. Barham, J. Chen, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (2016), pp. 265–283.
39. D. P. Kingma, "Adam: a method for stochastic optimization," in *Int Conf Learn Represent* (2014).
40. T. M. Khan, A. Robles-Kelly, and S. S. Naqvi, "T-Net: A resource-constrained tiny convolutional neural network for medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), pp. 644–653.
41. Y. Liu, J. Shen, L. Yang, *et al.*, "Wave-Net: A lightweight deep network for retinal vessel segmentation from fundus images," Comput. Biol. Med. **152**, 106341 (2023).
42. D. Su, J. Luo, and C. Fei, "An efficient and rapid medical image segmentation network," IEEE J. Biomed. Health Inform. **28**(5), 2979–2990 (2024).
43. Z. Gu, J. Cheng, H. Fu, *et al.*, "Ce-net: Context encoder network for 2d medical image segmentation," IEEE Trans. Med. Imaging **38**(10), 2281–2292 (2019).
44. H.-Y. Chou, S.-L. Huang, J.-W. Tjiu, *et al.*, "Dermal epidermal junction detection for full-field optical coherence tomography data of human skin by deep learning," Computerized Medical Imaging and Graphics **87**, 101833 (2021).
45. H. Ding, J. Q. Lu, W. A. Wooden, *et al.*, "Refractive indices of human skin tissues at eight wavelengths and estimated dispersion relations between 300 and 1600 nm," Phys. Med. Biol. **51**(6), 1479–1489 (2006).