Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Harmonious parameters and performance: Lightweight convolutional stage and local feature weighted fusion MLP for medical image segmentation ☆

Yan-Xu Chen, Yu-Jie Xiong *, Xi-He Qiu, Chun-Ming Xia *

School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

ARTICLE INFO

Keywords:
Medical image segmentation
Lightweight model
Local feature weighted fusion MLP
Large-scale kernel depth-wise separable convolution

ABSTRACT

Transformer-based models is widely applied in the field of medical image processing in recent years, thanks to its advantage in capturing global representations. However, while reaping high performance, the significant computational complexity, high training cost, and redundant dependencies of the model cannot be ignored. Therefore, our attention is drawn to the Multilayer Perceptron (MLP) as the complex self-attention alternative. We propose a lightweight image segmentation model based on depth-wise separable convolution and MLP, named **UConvNeXt**. It comprises both initial convolutional stages and MLP stages in the latent stage. Specifically, we employ large-scale kernel depth-wise separable convolution to replace the convolution blocks in UNet. The parameter is significantly reduced while maintaining the performance of the original model. In contrast to static feature fusion, we propose a novel local feature weighted fusion MLP (**LFWF-MLP**) module. In this module, contextual information is captured by shifting all tokens along the spatial directions. Feature correlations between different positions are weighted fused, and key regions receive increased attention, ultimately enhancing the segmentation performance. The experimental results demonstrate that, compared to models with similar parameter levels, more powerful segmentation performance is exhibited by UConvNeXt. Moreover, when compared to models with higher parameters than ours, comparable or even slightly superior segmentation results are achieved by UConvNeXt. In terms of model scale, compared to the UNet, parameters is reduced by 17 times, and computational complexity is lowered by 14 times, and inference speed is improved by 3 times in UConvNeXts.

1. Introduction

As the essential technique in the field of medical diagnosis and image analysis, medical image segmentation is utilized to accurately extract crucial information include anatomical structures, pathological tissues, or body organs from medical images. A better understanding of the disease are achieved through this process by doctors, researchers, and medical imaging professionals. For example, in the early diagnosis and treatment of melanoma, the suspicious lesions on the skin are segmented using medical image segmentation techniques, allowing potential melanoma areas to be better located and quantified by doctors. However, due to time-consuming and error-prone process, manually annotating these critical information in clinical work is not feasible. Therefore, there is a rapid growth in demand for automated and accurate lesions segmentation. With the rapid development of deep learning [1–3], an increasing number of deep neural networks have been proposed for medical image segmentation [4,5]. The powerful representation learning capabilities and end-to-end training approach of these models are leveraged by researchers, leading to significant advancements in the field of medical image processing.

Since 2015, models based on U-Net [4] have been dominating the development of medical image segmentation. The unique aspect of the U-shaped network is the symmetric encoder–decoder architecture [6]. In the encoder part, the image is subjected to several down-sampling operations, gradually reducing its size and extracting high-level features. Then, in the decoder part, the feature maps are up-sampled to the same size as the input image, and fine segmentation results are gradually reconstructed. After the proposal of Vision Transformer [7], the combination of it and the U-shaped architecture started to be explored by researchers [8,9]. Leveraging the advantages of Transformers in global modeling and capturing long-range dependencies, along with the

☆ This work was supported by the National Natural Science Foundation of China (62006150) and Science and Technology Commission of Shanghai Municipality, China (21DZ2203100).

* Corresponding author.
E-mail addresses: m325122625@sues.edu.cn (Y.-X. Chen), xiong@sues.edu.cn (Y.-J. Xiong), qiuxihe@sues.edu.cn (X.-H. Qiu), cmxia@sues.edu.cn (C.-M. Xia).

https://doi.org/10.1016/j.bspc.2024.106726
Received 10 October 2023; Received in revised form 16 June 2024; Accepted 1 August 2024
Available online 14 August 2024

inductive bias capabilities of CNN, further enhances the performance of medical segmentation models.

However, due to the unique self-attention in Transformer models, correlation computation is required between each token, resulting in high computational complexity. For the Transformer model with an input image size is n, the spatial complexity of its self-attention is as high as $O(n^2)$. The indispensability of the Transformer model and the positive contributions of each parameter in the self-attention to the final results are currently being questioned by many researchers [10–12]. Furthermore, the transition of segmentation techniques from laboratory settings to clinical environments (where testing and analysis are performed in the presence of patients) is hindered by computational latency and large graphics memory overhead.

In this work, we focus on addressing the aforementioned issues and propose a cost-effective lightweight model called UConvNeXt. It is designed to have fewer parameters and computational expenses while still maintaining the same segmentation performance as high-parameter models. Specifically, the design of UConvNeXt follows the principles of the U-shaped network and is divided into two parts: the convolutional stage and the MLP stage. In the convolutional stage, we use large-scale kernel depth-wise separable convolution to improve the convolution blocks of UNet. In the lower part of the model, we design a novel local feature weighted fusion MLP (**LFWF-MLP**) module. In the module, convolution features are projected into abstract tokens, and all tokens are moved along the spatial dimensions to capture contextual information. Feature correlations between different directions are computed and used as weights to guide feature fusion. We evaluate UConvNeXt on the ISIC-2018 dataset [13], the 2018 Data Science Bowl [14], and the Breast UltraSound Images (BUSI) [15]. The results show that Compared to models with similar parameters, UConvNeXt demonstrates stronger segmentation performance. When compared to high-parameter models such as TransUNet [5] (with 59 times more parameters and 9 times more analysis speed) and FAT-Net [16] (with 17 times more parameters and 5 times more analysis speed), UConvNeXt achieves comparable segmentation results.

The following are main contributions of this work:

- We propose a lightweight model named UConvNeXt for medical image segmentation. In the model, large-scale depthwise separable convolutions are employed for feature extraction in the initial convolutional stages. The LFWF-MLP module is utilized in the latent stage to fuse features. The former enables the model to achieve lightweight, while the latter enhances the segmentation capabilities of model.
- We propose the LFWF-MLP module to enhance the ability to fuse features between different pixels. Within the module, context information is captured by moving all tokens along the spatial dimension. Feature correlations between different tokens are weighted and fused, increasing the focus on critical regions and ultimately improving segmentation performance.
- We conduct experimental analysis using four public medical image segmentation datasets, including ISIC-2018 [13], 2018 Data Science Bowl [14], BUSI [15], and Synapse multi-organ segmentation dataset [17]. The experimental demonstrate that UConvNeXt achieves better results compared to the SOTA method. In terms of model scale, compared to the standard UNet, parameters is reduced by 17 times, and computational complexity is lowered by 14 times, and inference speed is improved by 3 times.

2. Related work

This section primarily introduces deep learning models for medical image segmentation, such as CNN-based and Transformer-based models, as well as the MLP model, which has recently been frequently used as an efficient alternative to transformer. It also elaborates on the characteristics and advantages of each model.

2.1. CNN-based models

The Fully Convolutional Network (FCN) [18] is significant milestone in the field of image segmentation. In FCN, traditional fully connected layers are replaced with convolution layers, and the image is pixel-wise mapped to segmentation results, preserving the spatial information of the image. In U-Net [4], an encoder–decoder architecture is employed, where feature information is extracted by the encoder through several layers of convolution and pooling operations, and restored to the original image size using up-sampling and convolution operations by the decoder. Recently, a pure convolutional neural network model named ConvNeXt was designed by Liu et al. [19]. In their work, the inverted bottleneck structure and large convolution kernels are reexamined, and deep convolution operations are employed in image classification tasks, resulting in performance surpassing the Swin Transformer [20]. This has once again sparked researchers' interest in the study of convolutional neural networks. In the work by Han et al. [21], the convolution blocks of UNet were improved using large convolution kernels and depth-wise separable convolution, leading to a significant reduction in the number of parameters. Furthermore, a lightweight attention mechanism is designed, where noise in low-level semantic information is filtered out, allowing more focus on the target region.

2.2. Transformer-based models

In recent research, Transformer-based models have gradually become mainstream due to their advantages in global feature modeling and the representation of global information. Chen et al. first combined Transformer with U-Net and proposed TransUNet [5]. The labeled image patches are encoded using the self-attention to extract global context, and precise localization is achieved by combining the high-resolution features extracted by CNN in the decoding stage. The first fully Transformer-based image segmentation model called Swin-Unet [22] was proposed by Hu Cao et al. Local-to-global self-attention is achieved in the encoder, and upsampling of global features to the input resolution is performed in the decoder for pixel-level segmentation predictions. UC-TransNet [9] was proposed by Haonan Wang et al. The model is comprised of a submodule that performs multi-scale channel cross-fusion and a submodule that guides the fusion of multi-scale channel cross-information. It is effectively connected to the decoder features, thereby eliminating ambiguity and accomplishing accurate segmentation.

2.3. In-depth analysis of transformer models

Hatamizadeh et al. proposed Swin UNITR [23], a model for brain tumor segmentation in MRI images, which replaces conventional U-Net encoding modules with hierarchical Swin Transformer blocks. Shifted window mechanism for self-attention are utilized by the blocks, effectively capturing long-range dependencies and multi-scale features. By combining Swin Transformer encoders and FCNN-based decoders, complex spatial relationships in medical images are simulated to achieve brain tumor segmentation. Zhou et al. proposed nnFormer [24], incorporating three attention mechanisms. Attention calculation within 3D local volumes is focused on by LV-MSA, with local dependencies being captured and computational complexity reduced. The feature pyramid with a wider receptive field is built by GV-MSA through the computation of self-attention on the global 3D volume. Integration between encoder and decoder information is enhanced by Skip Attention, which replaces traditional U-Net skip connections, leading to the recovery of finer prediction details. Shaker et al. proposed UNETR++ [25], a 3D medical image segmentation method emphasizing quality and efficiency. It introduces an Efficient Paired Attention module applying spatial and channel attention within two branches. Keys and values are projected into a low-dimensional space by spatial attention, enabling linear self-attention computation relative to input tokens. Parameter control is aided by shared weights between branches, balancing contributions to enhance performance.

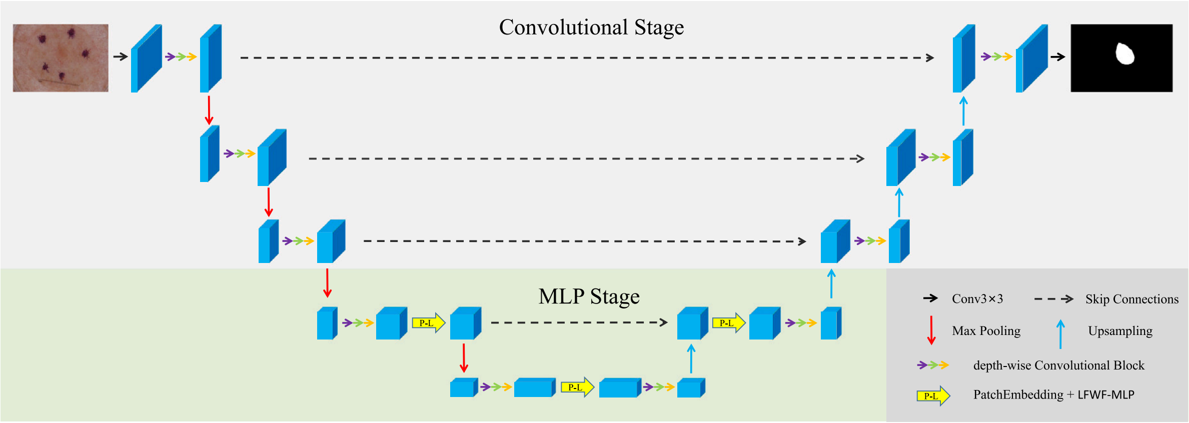


Fig. 1. UConvNeXt architecture diagram.

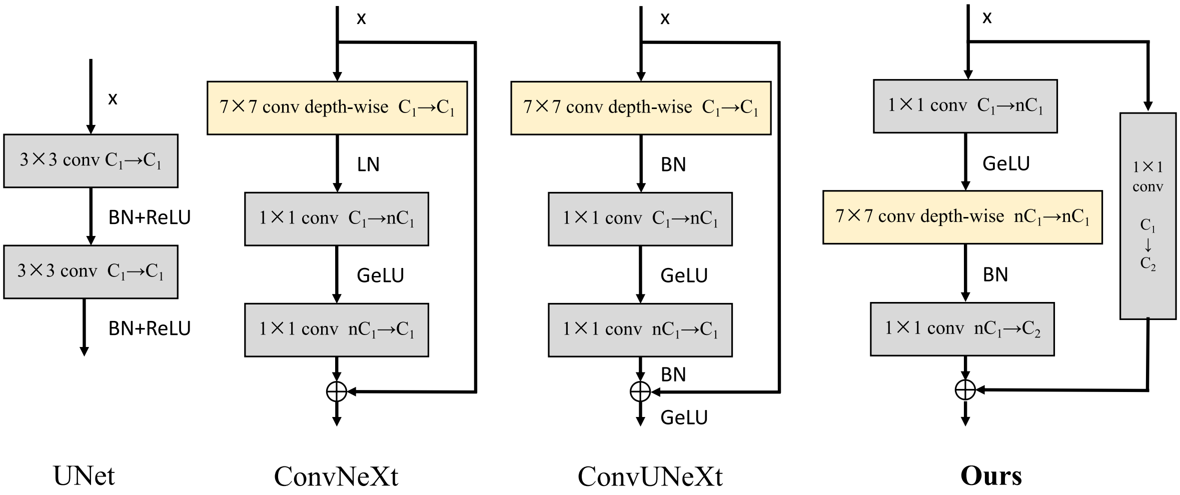


Fig. 2. Comparison chart of convolution blocks for various variants of UNet.

2.4. MLP-based models

A fully MLP-based model called MLP-Mixer [10] was designed by the Google. In this model, The encoding modules in ViT are replaced with token-mixing MLPs, and matrix transposition is creatively used to fuse channel and spatial information. By relying solely on basic matrix multiplication for data processing and feature extraction, performance similar to CNN and ViT models is achieved using a simple MLP model, nonlinear activation functions, and Layer normalization. Jeya et al. were the first to integrate MLP into the field of medical imaging and proposed a lightweight model called UNeXt [26]. Tokenized MLP blocks are utilized in the bottleneck to focus the MLP on local information of the convolution features in the model. In the tokenized MLP, the high-directional and wide-directional shift MLP operations are performed sequentially, followed by the incorporation of positional information through deep convolution. Our work combines this shift operation to extract features and establish feature connections between different tokens.

3. Method

The model architecture of UConvNeXt and the design principles of its internal modules are detailed in this section. As shown in Fig. 1, UConvNeXt can extract the corresponding lesion region $J \in \mathbb{R}^{1\times H\times W}$ from an input image $I \in \mathbb{R}^{3\times H\times W}$ of arbitrary size.

3.1. Convolutional stage

Depth-wise separable convolution. When encountering a resource-constrained environment, MobileNet [27] is highly favored, and its success owes much to the depth-wise separable convolution. The depth-wise separable convolution consists of two parts: depth-wise convolution and point-wise convolution. First, learnable depth-wise convolution kernels are used on each input channel. Then, the information from different channels is mixed through the point-wise convolution. As shown in Fig. 2, our design deviates from previous work and is more similar to the MobileNetV2 [28]. This is represented with the following formula:

$$X_1 = Conv_{1\times 1}(X, C_1, nC_1) \tag{1}$$

$$X_2 = depth\text{-}Conv_{7\times 7}(GELU(X_1), nC_1, nC_1) \tag{2}$$

$$Y = Conv_{1\times 1}(BN(X_2), nC_1, C_2) + Res(X) \tag{3}$$

where X_1, X_2, and Y represent the two intermediate features and the output feature, and C_1 and C_2 respectively denote the input and output channel numbers. Experimental results show that the parameters are significantly reduced while the accuracy does not change significantly.

Convolution kernel size. One of the key ideas in VGGNet [29] is the utilization of stacked consecutive 3 × 3 convolution layers instead of larger-sized kernels, which resulted in larger kernels being neglected for a period of time. However, in the Swin-Transformer, the effective

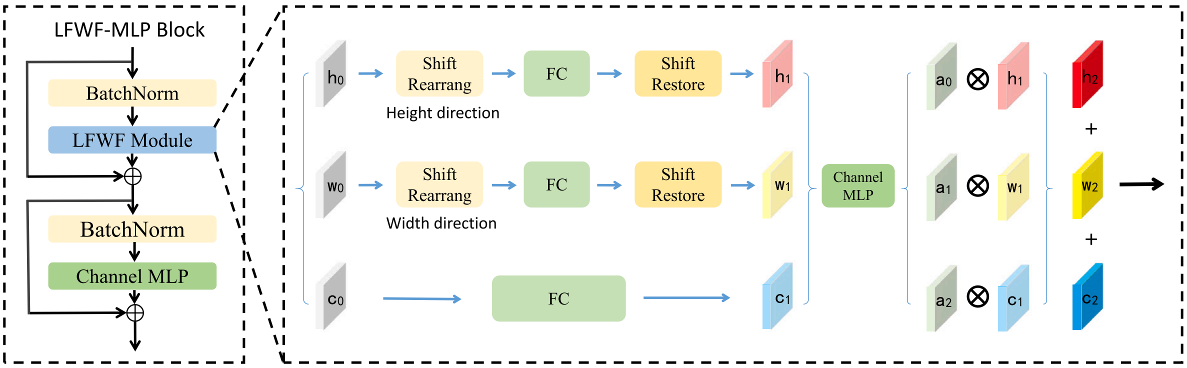


Fig. 3. The structure diagram of the local feature weighted fusion MLP (LFWF-MLP) is shown. The Rearrangement layer and Restoration layer in the LFWF-Model are illustrated in Fig. 4.

use of a 7 × 7 local window has prompted researchers to revisit the exploration of larger convolution kernels again [19,21]. We re-examined the reasons for the decrease in accuracy and found that the decrease in receptive field is the main obstacle. Several different sizes of convolution kernels are tried, including 3, 5, 7, and 9. We found that at a kernel size of 7, performance reached a saturation point. Due to the use of depth-wise separable convolution, the increase in kernel size does not lead to a significant increase in parameters. Moreover, the increase in receptive field while using larger kernels brings the segmentation performance close to that of the original model. Ultimately, a depth-wise separable convolution with a kernel size of 7 × 7 is selected.

Inverted bottleneck design. The issue of information loss during information flow propagation is creatively addressed in MobileNetV2 [28] by utilizing the Inverted bottleneck design. As shown in Fig. 2, UConvNeXt is designed similarly, where firstly the channel numbers are multiplied by a factor of n (n is a hyperparameter, and we set it to 2) through the first depth-wise convolution, and then the channel numbers are reduced back to the output channel size in the point-wise convolution.

Activation function and normalization layer. Unlike previous approaches using depth-wise separable convolution, the option of using fewer activation functions and normalization layers is chosen. As shown in Fig. 2, Only one activation function is set after the channel-wise expansion point-wise convolution to filter out irrelevant features introduced by the dimensionality increase. After the deep convolution fuses the features, a batch normalization layer is used to enhance the generalization ability of the channel-wise reduction point-wise convolution. Similar to ConvUNeXt, GELU [30] is employed as the activation function, and the use of BN [31] instead of LN [32] is experimented with.

Residual connection. Unlike ConvUNeXt, channel expansion is directly performed within the depth-wise separable convolution block. As a result, the inconsistency between input and output channel numbers prevents us from simply using path addition for residual connections. To address this, we introduce a point-wise convolution in the residual path to perform the transformation between input channel and output channel dimensions.

3.2. Local feature weighted fusion MLP

As shown in Fig. 3, the LFWF-MLP consists of two sub-blocks connected in series: the LFWF-module and the channel MLP. They respectively aggregate spatial information and channel information. Given an input feature $X \in \mathbb{R}^{H\times W\times C}$ with height H, width W, and channel count C, the LFWF-MLP block is formulated as follows:

$$Y = LFWF\text{-}Module(BN(X)) + X \tag{4}$$

$$Z = Channel\text{-}MLP(BN(Y)) + Y \tag{5}$$

where Y and Z represent the intermediate features and output features of the module, respectively. We place three LFWF-MLPs in the last two layers of the model to guide the fusion of deep features and suppress irrelevant features. Compared to Hire-MLP [33] and DynaMixer [34], the main difference is that we replace their Hire Module and DynaMixer Block with LFWF-module, which successfully captures the relationships between different tokens.

Shift MLP. With reference to previous works [26,35], The shift operation is considered meaningful for the integration of semantic information. Therefore, the feature maps extracted by deep separable convolution are projected into non-overlapping tokens and then subjected to batch normalization. Prior to the shift operation, the feature maps are divided into s segments along the channel dimension (s is a hyperparameter, and in our experiments, we set it to 5). In the shift operation, tokens are moved along the height and width directions to exchange information from different regions, thereby allowing the model to aggregate spatial information among features. As shown in Fig. 4, the operation shifts all tokens in a given direction with a given step size s. After the shift operation, the current region's features are aggregated with features that are $\lfloor \frac{S}{2} \rfloor$ positions apart, enabling communication between local features. Following the fully connected operation, the positions of the shifted tokens are restored to maintain the relative positions between different tokens.

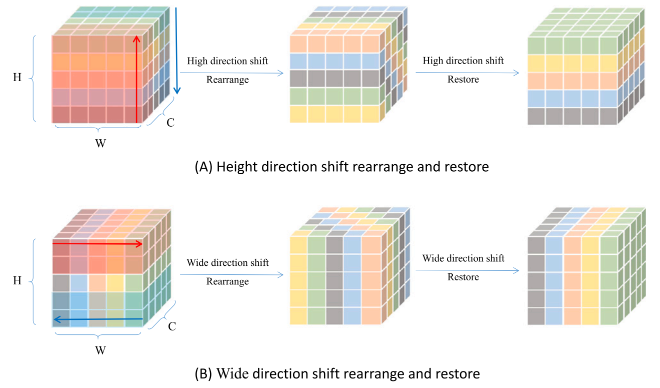
Unlike the UNeXt, we do not add padding to handle boundary feature information. We think it causes the loss of some useful features. Our operation is similar to the Cross-region Rearrangement [33], shifting the current boundary tokens to the opposite boundary. This ensures the integrity of deep features. If the features from the opposite boundaries are irrelevant to the current feature, they will be assigned a low weight during the weighted fusion process. Meanwhile, this operation reduce computational complexity.

LFWF-Modele. The input feature X with dimensions $H \times W \times C$, after undergoing batch normalization, is replicated into three copies. Spatial information communication occurs within two of these branches, namely along the height and width directions. Based on previous work [33,34], we also add an additional branch without spatial communication. The input X is sent to these three branches to respectively fuse local features from relative directions, resulting in the fused feature (h_1, w_1, c_1). The stacked features are then fed into a small-scale channel MLP. The output consists of the importance weights (a_0, a_1, a_2) for the three different directional features. The formulas are as follows:

$$(a_0, a_1, a_2) = FC_2(GELU(FC_1(h_1, w_1, c_1))) \tag{6}$$

$$F = h_1 \times a_0 + w_1 \times a_1 + c_1 \times a_2 \tag{7}$$

Finally, the three fused feature maps with different directional information, (a_0, a_1, a_2), are linearly combined. Among them, F serves as the final output of the LFWF Module, where the respective weights are calculated based on the three directional branches and the information



(A) Height direction shift rearrange and restore

(B) Wide direction shift rearrange and restore

Fig. 4. The explanation of the Shifting rearrangement operation in the LFWF-Model. Features are shifted in both height and width dimensions to capture contextual information.

between the current feature and the surrounding features is dynamically fused. Due to the presence of the weighted fusion mechanism, irrelevant features will be assigned extremely low weights, allowing UconvNeXt to focus more on target areas with high weights and high relevance.

It is worth mentioning that we only employ one fully connected layer in both the height and width directional branches. Experimental results show that using more fully connected layers, similar to the work in [33], does not result in higher segmentation accuracy during model testing. On the contrary, it leads to stronger over-fitting phenomena.

3.3. Overall architecture

As shown in Fig. 1, the overall model is divided into two parts: the convolutional stage and the MLP stage. We still follow the 5-layer encoder–decoder architecture of UNet, but with modifications in the design of each block. The first three layers of the model constitute the convolutional stage, where the original convolution blocks are replaced with large-scale kernel depth-wise separable convolution that consist of three parts. In the last two layers of the model, LFWF-MLP is added after the depth-wise separable convolution to fuse features from the convolutional stages, thereby obtaining high-level features with a large receptive field. To reduce the parameters, we employ 2×2 max-pooling layers for feature down-sampling and use linear interpolation for feature up-sampling. In the skip connections, we use addition instead of concatenation. Referring to previous work [26], we set the channel numbers of each layer as $(C_1, C_2, C_3, C_4, C_5) = (16, 32, 64, 128, 256)$ in our model.

3.4. Loss function

Medical image segmentation is viewed as a pixel-level binary classification task, where the input data needs to be divided into two mutually exclusive classes. In such cases, binary cross-entropy loss function is a good choice. It is expressed as follows:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[\hat{y}\log(y_p) + (1-\hat{y})\log(1-y_p)] \tag{8}$$

Among them, y_p is the ground truth, $\hat{y}$ is the predicted segmentation mask, and N is the number of pixels. In the field of image segmentation, the Dice coefficient is commonly used to measure the similarity between the predicted segmentation result and the ground truth [36,37]. By minimizing the Dice loss function, it promotes the generation of more accurate results. It is expressed as follows:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \times \sum_{i=1}^{N} y_p \hat{y}}{\sum_{i=1}^{N} y_p + \sum_{i=1}^{N} \hat{y}} \tag{9}$$

Table 1
Parameters of the ISIC-2018, BUSI, 2018-DSB and synapse datasets.

Dataset	Imaging	Shape	Images	Train	Test
ISIC-2018	Skin	512 × 512	2,596	2,076	520
2018-DSB	Nucleus	512 × 512	670	536	134
BUSI	Cancer	256 × 256	647	518	129
Synapse	Multi-organ	512 × 512	3779	2212	1567

To comprehensively consider both pixel-wise classification accuracy and segmentation precision, we utilize a combination of binary cross-entropy and Dice loss to train UConvNext. The final loss function is as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{BCE}(y_p, \hat{y}) + \lambda_2\mathcal{L}_{Dice}(y_p, \hat{y}) \tag{10}$$

In the formula, the weights of $\mathcal{L}_{Dice}$ and $\mathcal{L}_{BCE}$ are determined by previous experimental experience [26,35], with λ_1 set to 0.5 and λ_2 set to 1.

4. Experiments

In this section, the segmentation performance of UConvNeXt is evaluated on three different medical segmentation tasks, and it is compared with SOTA models. Then, ablation studies are conducted to analyze the effects of each module used in UConvNeXt.

4.1. Datasets

In order to make our experiments as clinically relevant as possible, we utilized three distinct types of public medical image datasets to evaluate our model, include International Skin Imaging Collaboration (ISIC-2018) [13], Breast UltraSound Images (BUSI) [15], the 2018 Data Science Bowl (2018-DSB) [14], and Synapse multi-organ segmentation dataset [17]. Detailed statistics are provided in Table 1.

- **ISIC-2018 Dataset:** This dataset is sourced from the ISIC-2018 Challenge [13,38] and has become a primary benchmark for evaluating skin lesion image segmentation algorithms. It consists of 2596 images with corresponding annotations. We resized the images to a resolution of 512 × 512. The dataset is randomly divided into a training set and a test set, with each set comprising 80% and 20% of the total dataset.
- **2018-DSB Dataset:** The dataset is sourced from the 2018 Data Science Bowl Challenge [14] and is used for locating cell nuclei in divergent images. It comprises 670 images with corresponding annotations, all resized to a resolution of 512 × 512.
- **BUSI Dataset:** The dataset consists of ultrasound images of normal, benign, and malignant cases of breast cancer, along with their corresponding segmentation masks. We only utilized the benign and malignant images, totaling 647 images, resized to a resolution of 256 × 256.
- **Synapse:** This dataset comprises 30 cases of abdominal CT scans, with each CT slice annotated for 13 organs. Each CT volume consists of 85 to 198 slices, resulting in a total of 3779 axial enhanced CT images of the abdomen. The images are of size 512 × 512 pixels, with voxel spatial resolutions ranging from $([0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0])$ mm^3.

4.2. Experimental settings

Due to the difficulty in collecting an adequate number of training samples in medical imaging, severe over-fitting is often experienced by models during training [39]. To alleviate this issue, we employ four random data augmentation operations during the model training phase, including random horizontal flipping, random vertical flipping, random

Table 2
Comparison with various SOTA models on ISIC-2018 dataset.

Method	Year	Params (in M)	GFLOPs	IoU (%)	Dice (%)	Acc (%)	Recall (%)	Precision (%)
UNet [4]	2015	31.13	55.84	80.2 ± 0.18	87.4 ± 0.15	95.2 ± 0.08	90.6 ± 0.18	88.3 ± 0.15
UNet++ [40]	2018	9.16	34.65	75.1 ± 0.17	84.9 ± 0.07	95.4 ± 0.08	90.6 ± 0.16	89.9 ± 0.13
ResUNet++ [41]	2019	87	94.56	82.2 ± 0.14	89.4 ± 0.11	95.4 ± 0.05	90.3 ± 0.10	90.0 ± 0.12
AttU-Net [42]	2019	34.88	72.81	82.19 ± 0.38	89.13 ± 0.31	95.56 ± 0.27	89.98 ± 0.17	91.50 ± 0.14
UNeXt [26]	2021	1.48	2.28	82.93 ± 0.16	90.21 ± 0.15	95.52 ± 0.09	91.70 ± 0.18	91.34 ± 0.12
APFormer [8]	2022	2.6	4.1	83.47 ± 0.40	90.07 ± 0.15	96.06 ± 0.10	90.99 ± 0.87	–
DCSAU-Net [39]	2023	2.6	6.91	84.10 ± 0.16	90.40 ± 0.13	96.10 ± 0.08	91.07 ± 0.12	92.20 ± 0.14
TransUNet [5]	2021	105.32	38.52	77.0 ± 0.20	84.9 ± 0.18	94.5 ± 0.09	89.8 ± 0.19	84.7 ± 0.19
FAT-Net [16]	2021	30	23	82.02	89.03	95.78	91.00	–
MSCA-Net [43]	2023	27.09	12.88	**84.18 ± 0.38**	90.52 ± 0.26	96.41 ± 0.29	–	–
ACCPG-Net [44]	2023	11.8	–	83.52	90.81	96.13	88.61	**93.82**
TransAttUNet-D [45]	2023	25.96	67.69	83.80	90.74	96.38	90.93	92.42
DAE-Former [46]	2023	49.21	26.50	83.10 ± 0.34	90.50 ± 0.29	95.99 ± 0.21	91.92 ± 0.11	91.51 ± 0.22
Ours	–	1.76	2.44	83.93 ± 0.11	**90.89 ± 0.11**	**96.62 ± 0.07**	**92.84 ± 0.16**	91.64 ± 0.12

cropping, and random rotation by a certain angle within the range of $(-\pi/2, \pi/2)$.

All experiments are conducted using the PyTorch framework with an NVIDIA GeForce RTX 2080Ti graphics card and 8 GB of memory. To ensure the reproducibility, we maintain the same experimental settings as previous researchers [26]. The Adam optimizer with a learning rate of 0.001 and a momentum of 0.9 is employed. Additionally, a cosine annealing learning rate scheduler with a minimum learning rate of 0.00001 is utilized. The batch size is set to 8. UConvNeXt is trained for a total of 400 epochs. To ensure fairness in the experiments and following previous data partitioning practices, three 80 - 20 random splits in the dataset are performed (with $random\_state$ set to 0, 1, 2), and the model with the highest sum of Intersection over Union (IoU) and Dice coefficient on the validation set is saved. The mean and variance of these metrics are reported to analyze the model's performance [47,48].

4.3. Evaluation criteria

We adopt five widely used metrics in the field of image segmentation to evaluate the accuracy performance of our proposed models, including Intersection over Union (IoU), Dice coefficient, Accuracy (ACC), Recall, and Precision. The formulas for these evaluation metrics are as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{12}$$

$$ACC = \frac{TP + FN}{TP + TN + FP + FN} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

where True Positive (TP) represents the pixels correctly classified as the key region. False Positive (FP) represents the pixels incorrectly classified as the key region when they are actually non-key regions. Similarly, True Negative (TN) represents the pixels correctly classified as non-key regions, and False Negative (FN) represents the pixels that are not extracted as key regions.

4.4. Comparative results on ISIC-2018 dataset

Quantitative Analysis. Table 2 shows the performance comparison of UConvNeXt with thirteen SOTA models on the ISIC-2018 dataset. This includes four commonly used models (UNet [4], UNet++ [40], AttU-Net [42], ResUNet++ [41]), three models with similar parameter levels (UNeXt [26], DCSAU-Net [39], APFormer [8]), and six models with higher parameter levels (TransUNet [5], FAT-Net [16], MSCA-Net [43], ACCPC-Net [44], TransAttUNet [45], DAE-Former [46]). To ensure a fair comparison, we set up the same runtime environment as UNeXt, and the data for other models are sourced from their respective papers. First, comparisons are made with models of similar parameter levels. UConvNeXt achieves the best performance in most metrics. When compared with models of higher parameters, UConvNeXt demonstrates comparable performance and even exhibits a slight advantage.

Qualitative Analysis. Five representative models, including U-Net [4], U-Net++ [40], AttU-Net [41], TransUNet [5], and UNeXt [26], are selected for visual comparison. As shown in Fig. 5, when using early models like UNet and UNet++, many critical regions cannot be accurately delineated. In Att-UNet, an attention mechanism is added to the skip connections, allowing for more refined processing of details, but it may also result in misclassifications in many non-critical regions. TransUNet achieves good visual effects through global modeling. Our model demonstrates superior performance in handling local details compared to UNeXt.

4.5. Comparative results on 2018 DSB dataset

Quantitative Analysis. For the 2018 DSB dataset, UConvNeXt is compared with thirteen advanced models, including four commonly used models, three models with similar parameter levels, and six models with higher parameter levels. The comparison result is shown in Table 3. Compared to U-Net, UConvNeXt achieves improvements of 2.74% in IoU, 1.47% in Dice, 2.45% in Recall, and 1.24% in Precision. Meanwhile, parameters is reduced by 17 times, and computational complexity is lowered by 14 times. It is evident that UConvNext surpasses some SOTA methods.

Qualitative Analysis. A visual comparison is also conducted using the same five representative models as the ISIC-2018 dataset. As shown in Fig. 6, superior accuracy in cell nucleus segmentation is achieved by UConvNeXt compared to the other five competitors. Even for cells with different background colors and irregular shapes, The extracted contours by UConvNeXt are observed to closely resemble the ground truth.

4.6. Comparative results on BUSI dataset

Quantitative Analysis. For the BUSI dataset, UConvNeXt is compared with nine advanced models, including three commonly used models, two models with similar parameter levels, and four models with higher parameter levels. The comparison result is shown in

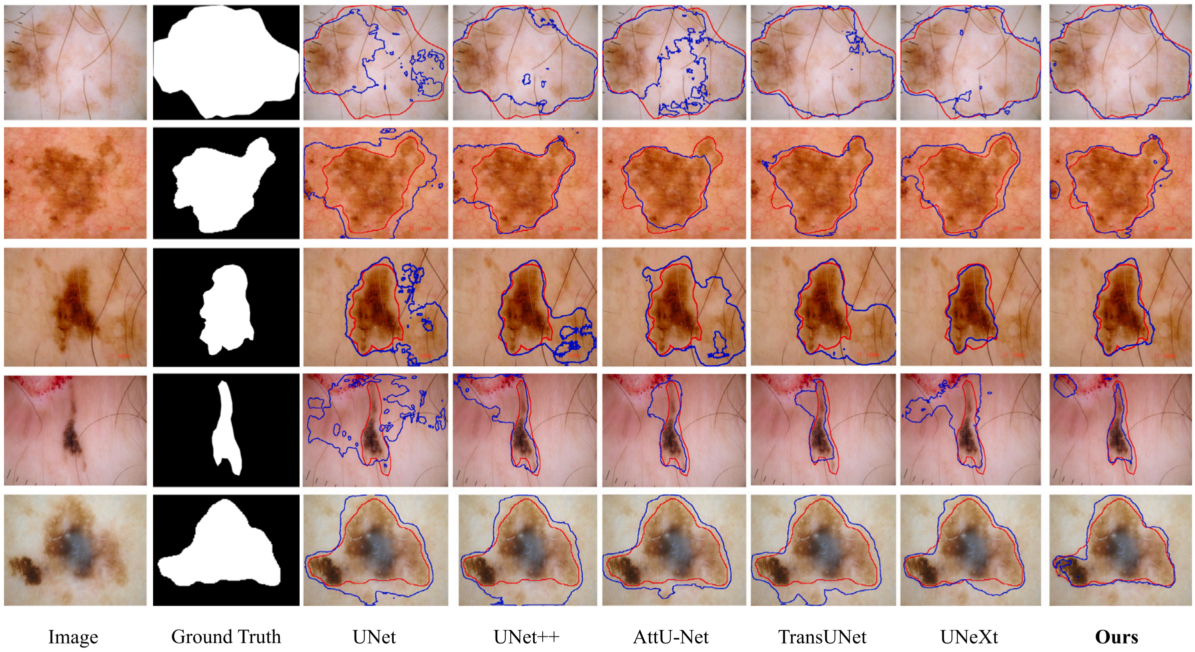


Fig. 5. Visual comparisons are conducted with various models on the ISIC-2018 dataset. The red contour represents the ground truth, while the blue contour indicates the segmentation results of the respective models.

Table 3
Comparison with various SOTA models on the 2018 DSB.

Method	Params (in M)	IoU (%)	Dice (%)	Recall (%)	Precision (%)
UNet [4]	31.13	83.14 ± 0.10	90.80 ± 0.06	90.29 ± 0.09	91.30 ± 0.07
UNet++ [40]	9.04	52.65 ± 0.30	77.05 ± 0.30	71.59 ± 0.31	66.57 ± 0.27
ResUNet++ [41]	4.07	83.70 ± 0.11	90.98 ± 0.08	91.69 ± 0.10	90.57 ± 0.08
AttU-Net [42]	34.88	85.70	91.79	91.83	92.35
FANet [49]	7.72	85.69	91.76	92.22	91.94
DDANet [50]	6.83	84.52 ± 0.11	91.82 ± 0.07	91.39 ± 0.02	**92.89 ±0.06**
UNeXt [26]	1.48	84.78 ± 0.16	91.57 ± 0.15	92.10 ± 0.18	91.11 ± 0.12
DoubleUNet [51]	29.29	84.29 ± 0.11	91.09 ± 0.09	92.78 ± 0.10	90.20 ± 0.10
TransUNet [5]	105.32	84.50 ± 0.33	91.50 ± 0.31	92.30 ± 0.17	89.7 ± 0.14
UACANet-L [52]	69.15	77.91 ± 0.13	86.88 ± 0.10	90.61 ± 0.11	84.14 ± 0.12
MSRF-Net [53]	18.38	85.34 ± 0.08	92.24 ± 0.05	**94.02 ± 0.07**	90.22 ± 0.06
TransAttUNet-D [45]	25.96	84.98	91.62	91.85	91.93
DAE-Former [46]	49.21	85.67 ± 0.12	92.07 ± 0.09	92.86 ± 0.22	92.57 ± 0.21
Ours	1.76	**85.88 ± 0.14**	**92.27 ± 0.13**	92.74 ± 0.11	92.54 ± 0.14

Table 4
Comparison with various SOTA models on the BUSI dataset.

Method	Params (in M)	IoU (%)	Dice (%)	Recall (%)	Precision (%)
UNet [4]	31.13	65.36 ± 1.81	74.35 ± 1.72	77.86 ± 2.24	78.30 ± 1.98
UNet++ [40]	9.04	61.38 ± 1.73	71.58 ± 2.09	71.44 ± 2.77	79.68 ± 3.07
AttU-Net [42]	34.88	57.09 ± 1.22	67.99 ± 1.18	66.97 ± 4.08	78.78 ± 4.67
SKNet [54]	3.94	68.10 ± 1.63	77.51 ± 0.68	79.53 ± 1.93	78.62 ± 1.66
UNeXt [26]	1.48	69.07 ± 0.42	79.67 ± 0.37	79.47 ± 0.23	83.61 ± 0.88
TransUNet [5]	105.32	**72.40 ± 1.86**	**80.97 ±1.67**	80.28 ±2.28	84.04 ±2.11
AAU-Net [55]	29.65	68.82 ± 0.44	77.51 ± 0.68	**81.10 ± 0.52**	79.61 ± 1.07
TransAttUNet-D [45]	25.96	69.87 ± 0.59	79.53 ± 0.42	80.62 ± 0.32	82.97 ± 0.40
DAE-Former [46]	49.21	68.87 ± 0.32	77.97 ± 0.39	78.82 ± 0.32	83.57 ± 0.28
Ours	1.76	70.21 ± 0.39	80.33 ± 0.35	80.01 ± 0.42	**84.20 ± 0.44**

Table 4. Compared to the models with similar parameters, UNeXt, improvements of 1.14%, 0.66%, and 0.59% in IoU, Dice, and Precision. On this dataset, our model is not performing as exceptionally well as it did on the previous two datasets. we think that the large-scale kernel depth-wise separable convolution is better suited for processing large-sized images. However, when dealing with small-sized images, the effectiveness of the large receptive field is weakened due to the smaller deep feature maps, thereby impacting performance. When TransUNet handles small-size images, each token has richer semantic information, so we are slightly inferior to it on this dataset.

Qualitative Analysis. Visual comparisons are also conducted using the same five models as the previous two datasets. As shown in Fig. 7, UNet and UNet++ roughly segment the lesion area, but some crucial information is missed. AttU-Net and TransUNet models have an advantage in global processing due to their attention mechanisms. However, our model outperforms them in handling local details. Even for tumors

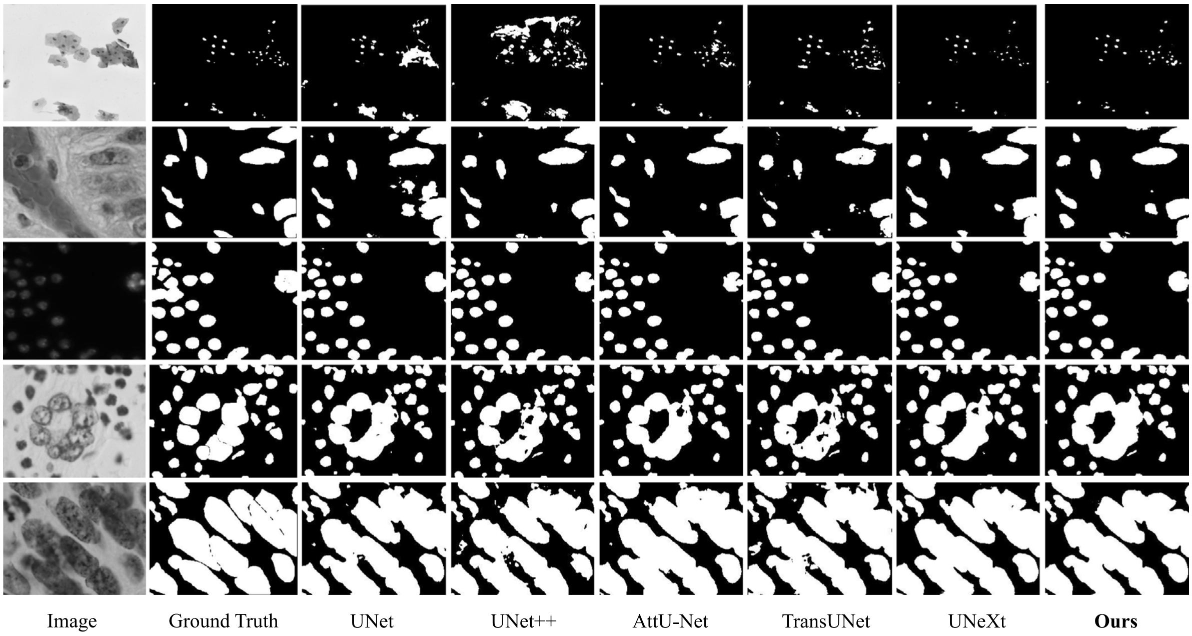


Fig. 6. Visual comparisons are conducted with different models on the 2018 DSB dataset.

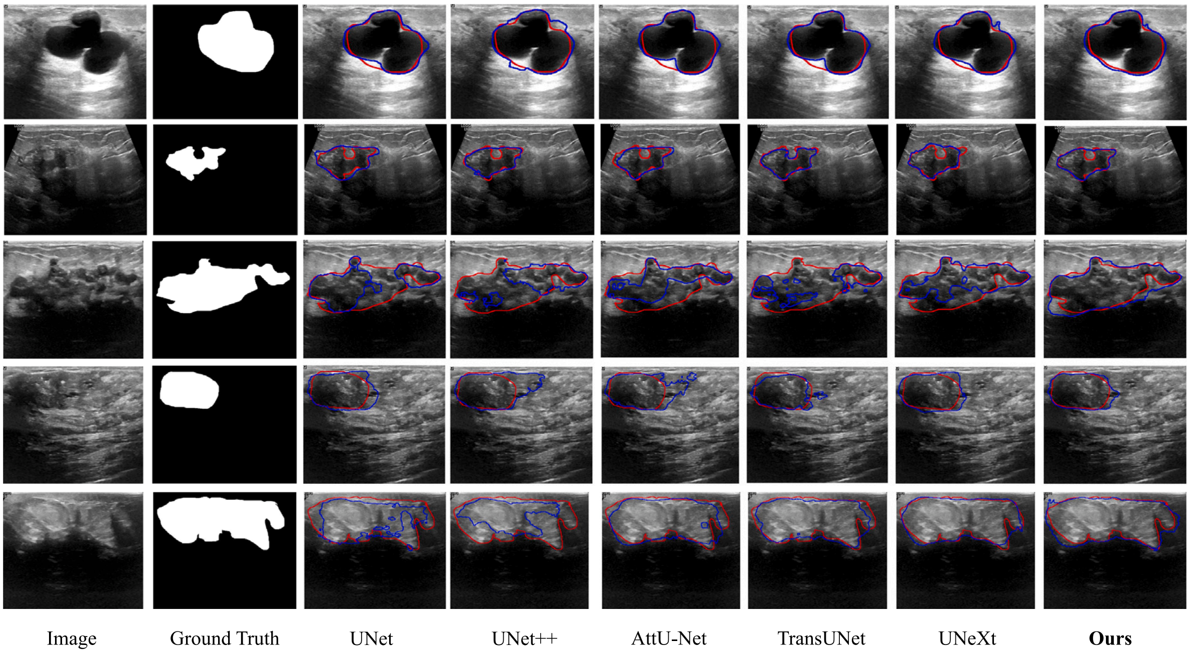


Fig. 7. Visual comparisons are conducted with different models on the BUSI dataset.

with varying scales, irregular shapes, and blurry boundaries, our results closely resemble the ground truth.

4.7. Comparative results on Synapse dataset

For the Synapse dataset, following the experimental setup of TransUNet, eight abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach) are utilized for experimental evaluation. In contrast to the previous three datasets, besides the Dice Coefficient, we employ the HD_{95} to assess the edge information of the segmented images. The HD_{95} metric calculates the distance between the segmented edge and the ground truth edge, and selects the $95th$ percentile of the distance after sorting to mitigate the influence of outliers, thus making the evaluation more robust. The formula is as follows:

$$HD_{95}(A,B) = MAX\{P_{95}(D(A,B)), P_{95}(D(A,B))\} \tag{16}$$

where, $D(A,B)$ represents the set of distances from each point in set A to the nearest point in set B, and $P_{95}(X)$ denotes the value in set X that is at the 95th percentile after sorting the values in ascending order.

Quantitative Analysis. UConvNeXt is compared with ten advanced models, including four traditional models (UNet [4], R50 U-Net, AttU-Net [42], TransUNet [5]), three lightweight models (UNeXt [26], CASTformer [56], APFormer [8]), and also selects three advanced 3D segmentation models (Swin-UNETR [23], nnFormer [24], UNETR++ [25]). The summarized quantitative comparison results are

Table 5
Comparison with various SOTA models on the synapse multi-organ segmentation dataset. (Abbreviations stand for: Aor: aorta, Gal: gallbladder, L-Kid: left kidney, R-Kid: right kidney, Liv: liver, Pan: pancreas, Spl: spleen, Sto: stomach.)

Type	Method	P(M)	GFLOPs	Organs								Average	
				Aor.	Gal.	L-Kid.	R-Kid.	Liv.	Pan.	Spl.	Sto.	Dice↑	HD95↓
Traditional model	U-Net [4]	31.13	55.84	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58	76.85	44.69
	R50 U-Net	144	34.65	84.18	82.84	79.19	71.29	93.35	48.23	84.41	73.92	74.68	36.87
	AttU-Net [42]	34.88	72.81	89.55	68.88	77.98	71.11	93.57	58.04	83.30	75.75	77.77	34.47
	TransUNet [5]	105.32	38.52	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	77.48	31.69
3D model	Swin-UNETR [23]	62.83	384.2	91.12	66.54	86.99	86.26	95.72	68.8	95.37	77.01	73.48	10.55
	nnFormer [24]	150.5	213.4	92.04	70.17	86.57	86.25	96.84	83.35	90.51	86.8	86.57	10.63
	UNETR++ [25]	42.96	47.98	92.52	71.25	87.54	87.18	96.42	81.1	95.77	86.01	87.22	7.53
Lightweight	UNeXt [26]	1.48	2.28	89.72	65.51	86.82	85.21	93.99	72.1	92.21	82.26	83.47	18.61
	CASTformer [56]	–	–	89.05	67.48	86.05	82.17	95.61	67.49	91	81.55	82.55	22.73
	APFormer [8]	2.6	3.9	90.84	64.36	90.54	85.99	94.93	72.16	91.88	77.55	83.53	16.37
	Ours	1.76	2.44	91.03	66.09	87.36	86.03	95.63	71.01	92.09	82.39	84.04	15.93

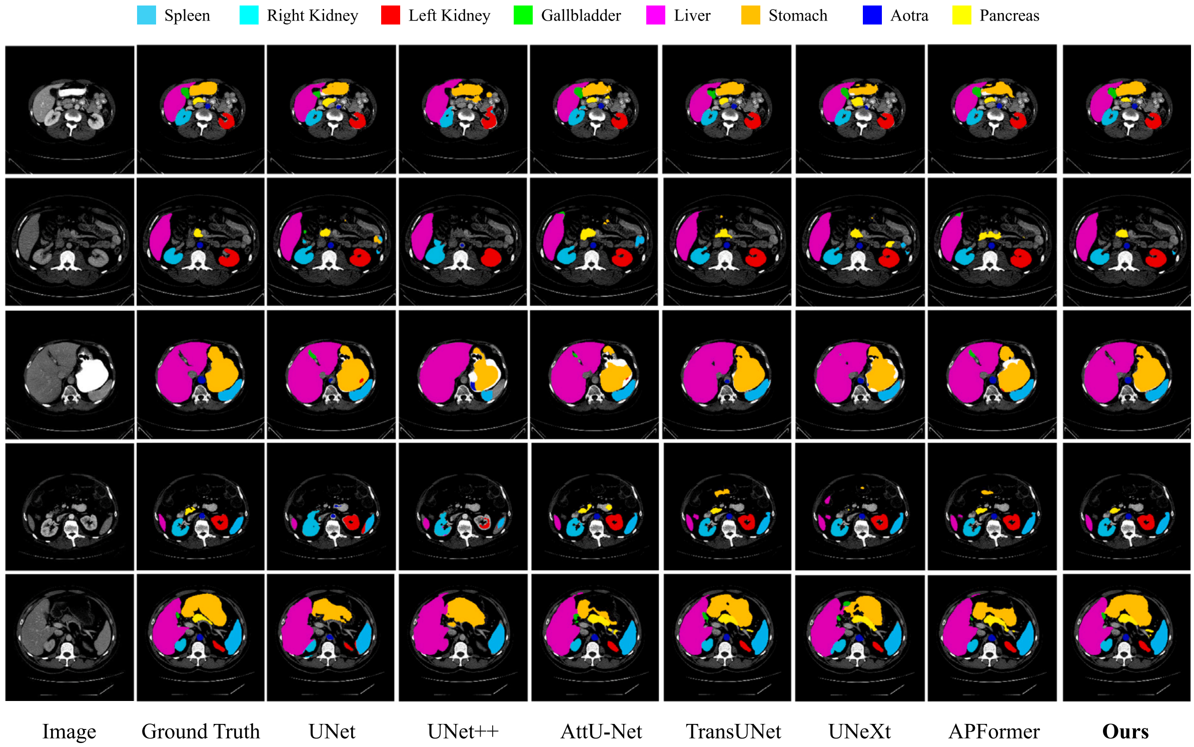


Fig. 8. Visual comparisons are conducted with different models on the Synapse multi-organ segmentation dataset.

shown in Table 5. Compared to traditional models, UConvNeXt demonstrates stronger modeling capability due to the presence of local feature weighted fusion, leading to better results. While 3D models excel in obtaining features in the horizontal and vertical directions of the plane images and simultaneously capture depth direction features of different planes, they incur computational costs that are ten times or even hundreds of times higher than lightweight models. In contrast, 2D-based methods can achieve satisfactory performance at lower complexity. As shown in the table, we outperform some 3D models in segmenting certain organs. Compared to lightweight models, we achieve leadership in most organ segmentation tasks and also outperform in average Dice and HD95. Specifically, compared to UNeXt with the same parameter size, Dice improves by 1.08%, while HD95 decreases by 3.68 mm. Compared to UNETR++, UConvNeXt is 25 times lighter in parameters and reduces computational complexity by 20 times, significantly enhancing the feasibility of clinical deployment.

Qualitative Analysis. The qualitative segmentation results of the six models on the synaptic dataset are shown in Fig. 8. As analyzed in APFormer, traditional convolutional models often exhibit segmentation errors in locally similar regions due to their inherent inductive biases. Transformer-based methods overcome this issue by modeling long-term dependencies. However, redundant dependencies adversely affect segmentation performance, potentially leading to misclassification in many non-critical areas. Conversely, UConvNeXt advantageously fuse local features by weighted integration, thus improving segmentation results significantly, particularly in handling local details.

4.8. Ablation analysis

As shown in Table 6, the ablation analysis is conducted from three different perspectives. The contribution of each module in UConvNeXt is demonstrated in the first group of experiments. The rationality of the design of the two modules is validated through the following two group of experiments. Experiments are conducted on three datasets, with the following analysis primarily based on the ISIC-2018 dataset, while others are provided for reference.

In the first group of experiments, We user a U-shaped model, which retains only 3 × 3 convolution compared to our model, as the baseline. As shown in Table 6, when using only the baseline, the IoU reached 79.71% and the Dice reached 86.91%. In the second row, the

Table 6
Ablation analysis for UConvNeXt on the ISIC-2018, 2018 DSB, BUSI datasets. In the first part of the table, the roles of different modules within the model are verified. In the second and third sections, the rationale behind the module designs is demonstrated.

Module	Method	Params (in M)	ISIC-2018		2018 DSB		BUSI	
			IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)
Architecture	Conv stage (**Baseline**)	1.19	79.71 ± 0.18	86.91 ± 0.16	82.42 ± 0.19	90.17 ± 0.15	64.88 ± 0.37	74.28 ± 0.27
	Conv → 7 × 7 DS Conv	0.68	79.49 ± 0.16	86.61 ± 0.13	81.97 ± 0.31	89.86 ± 0.21	64.47 ± 0.32	73.81 ± 0.21
	Baseline + LFWF MLP	2.38	83.81 ± 0.12	90.82 ± 0.12	85.84 ± 0.28	92.16 ± 0.21	70.07 ± 0.31	79.99 ± 0.27
	Conv → 7 × 7 DS Conv + LFWF MLP (**Ours**)	1.76	**83.93** ± **0.11**	**90.89** ± **0.11**	**85.88** ± **0.14**	**92.27** ± **0.13**	**70.21** ± **0.39**	**80.33** ± **0.35**
DS Conv	3 × 3 DS Conv	1.68	83.36 ± 0.12	90.45 ± 0.10	85.37 ± 0.23	91.79 ± 0.21	69.47 ± 0.31	79.67+0.21
	5 × 5 DS Conv	1.71	83.52 ± 0.15	90.47 ± 0.13	85.51 ± 0.17	92.11 ± 0.13	70.04 ± 0.34	80.19 ± 0.29
	9 × 9 DS Conv	1.83	83.57 ± 0.15	90.71 ± 0.14	85.45 ± 0.25	92.01 ± 0.21	69.49 ± 0.29	79.38 ± 0.21
	7 × 7 DS Conv no residuals	1.62	83.21 ± 0.15	90.14 ± 0.10	83.91 ± 0.26	91.45 ± 0.33	68.97 ± 0.21	78.97 ± 0.15
LFWF-MLP	LFWF-module direction(3→2)	1.64	83.12 ± 0.14	90.14 ± 0.11	85.01 ± 0.18	91.55 ± 0.16	68.97 ± 0.32	79.32 ± 0.27
	LFWF-module FC(1→2)	1.87	83.46 ± 0.18	90.59 ± 0.14	86.00 ± 0.28	91.77 ± 0.24	70.21 ± 0.25	80.21 ± 0.20
	LFWF-module without shifting	1.76	82.94 ± 0.10	89.77 ± 0.12	84.82 ± 0.29	91.19 ± 0.23	68.90 ± 0.29	78.82 ± 0.29
	LFWF-module without shift restore	1.76	83.49 ± 0.12	90.42 ± 0.10	85.38 ± 0.22	91.59 ± 0.20	69.62 ± 0.22	79.67 ± 0.18
	LFWF-module without weighted fusion	1.63	82.88 ± 0.15	89.68 ± 0.13	84.87 ± 0.21	91.23 ± 0.21	68.90 ± 0.23	78.72 ± 0.20

Table 7
Analysis on the number of channels.

Method	(C1, C2, C3, C4, C5)	Params (in M)	GFLOPs	ISIC-2018		2018-DSB		BUSI	
				Dice	IoU	Dice	IoU	Dice	IoU
UConvNeXt-S	(8,16,32,64,128)	0.48	0.81	89.21	82.54	91.15	84.01	78.37	67.98
UConvNeXt	(16,32,64,128,256)	1.76	2.44	90.89	83.93	92.27	85.88	80.33	70.21
UConvNeXt-L	(32,64,128,256,512)	6.44	7.61	91.21	84.24	92.48	86.21	81.31	71.03
UConvNeXt-XL	(64,128,256,512,1024)	25.37	27.01	91.39	84.47	82.62	86.39	81.77	71.48

convolution block is replaced with a large-scale kernel depth-wise separable convolution block, which significantly reduced the parameters but slightly decreased the segmentation performance. In the third row, the LFWF-MLP module is added to the baseline, resulting in a twofold increase in parameters and a significant improvement in segmentation performance. In the fourth row, which represents our final model, there is an improvement of 4.22% in IoU and an increase of 3.95% in the Dice compared to the baseline. It is worth mentioning that when the convolution block is simultaneously modified and the LEWF-MLP module is added, not only is there a decrease of 0.62M parameters, but there is also a slight improvement in performance. We think it is because the features extracted by the three-layer depth separable convolution block have a larger receptive field, making them more suitable for subsequent feature fusion with LFWF-MLP.

In the second group of experiments, the rationality of using large-scale kernel depth-wise separable convolution is validated, and the impact of removing residual connections is examined. As shown in the middle section of Table 6, We adjusted the kernel size of depth-wise separable convolution to 3 × 3. Compared to using standard convolution, there is a slight decrease of 0.33% in IoU and a slight decrease of 0.27% in Dice. The segmentation performance reaches its peak when the kernel size is increased to 7 × 7, but when further increased to 9 × 9, the performance decrease. It is speculated that an excessively large receptive field hinders the learning of local features in the lower layers. Regarding the use of residual connections, when it is removed, the performance decreases by 0.64% in IoU and 0.66% in Dice.

In the third group of experiments, we made five comparative modifications to LFWF-MLP. In the first row of the last section of Table 6, we removed the additional branch lacking spatial communication in the LFWF-Model, reducing the three branches to two, resulting in a decrease of approximately 0.60%. In the second row, following the traditional design approach [33,34], we initially employed two fully connected layers in each branch of the LFWF-Model. However, we find that using just one layer yields better segmentation performance. We analyze that the local correlations between tokens are effectively fused through a single fully connected layer following the shift operation. Further applying an additional fully connected layer disrupts the already fused features, leading to model overfitting. In the third and fourth rows, experimental results are presented to demonstrate the effects of not using shift operations and using shift operations without shift restore in the LFWF-MLP, aiming to validate the role of shifting. In the last row, we switched from using weighted fusion for the three directions to a simple additive approach, resulting in a nearly 1% decrease in overall model performance.

Analysis on number of channels. The number of channels in each layer of UConvNeXt is one of the most important hyperparameters, affecting the operating efficiency and segmentation performance. In Table 7, we perform a fourth set of ablation experiments on three datasets, showing three additional models of different scales. It is observed that when we double the number of channels (UConvNeXt-L), the segmentation performance is further improved, accompanied by an increase in computational overhead. When we reduce the number of channels to half (UConvNeXt-S), although the segmentation performance is reduced (reduced and Not drastic), but we can harvest a very lightweight model. When we upgrade the model to the same size as the standard UNet (UConvNeXt-XL), the model obtains richer image features due to the increase in channels. Currently, the performance of the model surpasses that of state-of-the-art methods.

4.9. Inference speed and model scale

We selected 500 images from the ISIC-2018 test dataset to evaluate the average speed of model inference on single images. In Fig. 9, We draw a comparison chart between model inference speed and the Dice coefficient, where the area of each bubble is proportional to the model's parameters. We have a model inference speed that surpasses most models (only slightly behind UNeXt), while simultaneously achieving the best segmentation performance. Compared to the typical UNet, the number of parameters is reduced by 17 times, and inference speed is improved by 3 times in UConvNeXt. The models parameter be found in Table 2.

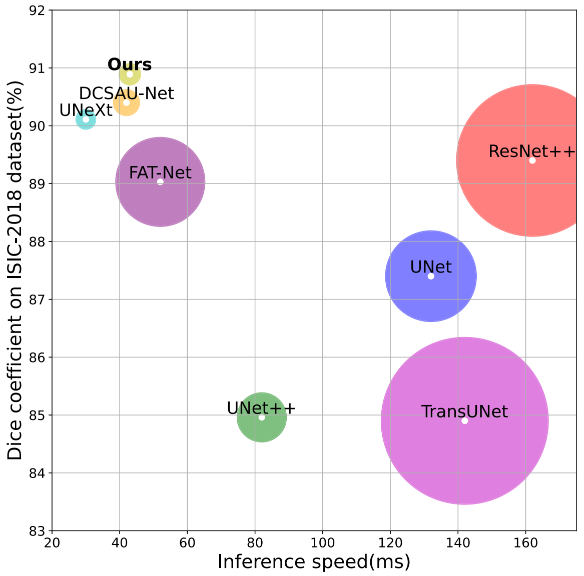


Fig. 9. Comparison chart of inference time and segmentation performance, where the area of each bubble is proportional to the model's parameters.

4.10. Limitation analysis

Suitable for plane images. The model is primarily designed for 2D plane images, so when dealing with 3D datasets, the volume needs to be segmented into a series of slices for processing. It renders the capturing of depth-oriented features challenging, resulting in suboptimal outcomes. This limitation stems from the fundamental characteristics of the model design, namely the lack of effective utilization of 3D feature in the data. To overcome the limitations of the model in handling 3D data, our future work can involve introducing more complex architectures or employing models specifically designed for 3d data, such as Point Transformer [57], etc. Additionally, exploring new data representation methods could allow better input of 3D datasets into 2D models without losing depth information [58]. Furthermore, employing multi-scale or hierarchical approaches could comprehensively capture image information across different dimensions.

Suitable for large size images. The model adopts depth-separable convolution with large convolution kernel, which shows excellent performance when processing large-size pathology images. However, when faced with small-size images such as 256 × 256, its performance is not as ideal as expected. The root cause of this limitation is that the model has undergone four downsampling processes, reducing the 256 × 256 image to the size of 16 × 16. It is worth noting that when operating on a 16 × 16 feature, the large-size convolution kernel faces challenges such as information loss and spatial discontinuity, which affects the segmentation effect of the model. In order to solve this problem, we will try to enhance the adaptability of the model to small-size images by introducing a more flexible size adaptation mechanism in future work, such as using smaller-size convolution kernels on small-size images. In addition, multi-scale skip connections can also be introduced to improve the model's generalization ability for images of different sizes.

5. Conclusion

In this work, we propose a novel lightweight medical image segmentation model called UConvNeXt. It is based on depth-wise separable convolution and MLP architecture, where we improve the convolutional stages by utilizing deep separable convolution with large kernel sizes, significantly reducing the parameters. In the MLP stages, we propose the LFWF-MLP module, which dynamically fuses the current pixel features with surrounding pixel features using fewer parameters, resulting in better segmentation performance by capturing key information. We validate UConvNeXt on three datasets and demonstrate its superior segmentation performance compared to models with similar parameter levels. Even when compared to some SOTA with significantly higher parameters, our model achieves comparable segmentation results.

CRediT authorship contribution statement

Yan-Xu Chen: Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Yu-Jie Xiong:** Visualization, Validation, Software, Investigation, Funding acquisition, Formal analysis. **Xi-He Qiu:** Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Chun-Ming Xia:** Supervision, Resources, Project administration, Funding acquisition, Formal analysis.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 30–39, https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[3] Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105, http://dx.doi.org/10.1145/3065386.

[4] Olaf Ronneberger, Fischer Philipp, Brox Thomas, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, et al., Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[6] Geoffrey E. Hinton, Osindero Simon, Yee-Whye Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554, http://dx.doi.org/10.1162/neco.2006.18.7.1527.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[8] Xian Lin, Li Yu, Kwang-Ting Cheng, Zengqiang Yan, The lighter the better: Rethinking Transformers in medical image segmentation through adaptive pruning, IEEE Trans. Med. Imaging 42 (7) (2023) 2325–2337, http://dx.doi.org/10.1109/TMI.2023.3247814.

[9] Haonan Wang, Peng Cao, Jiaqi Wang, Osmar Zaiane, Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer, Proc. AAAI Conf. Artif. Intell. 36 (3) (2022) 2441–2449, http://dx.doi.org/10.1609/aaai.v36i3.20144.

[10] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Mlp-mixer: An all-mlp architecture for vision, Adv. Neural Inf. Process. Syst. 34 (2021) 24261–24272, https://proceedings.neurips.cc/paper/2021/hash/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Abstract.html.

[11] Hanxiao Liu, Zihang Dai, David So, Quoc Le, Pay attention to mlps, Adv. Neural Inf. Process. Syst. 34 (2021) 9204–9215, https://proceedings.neurips.cc/paper/2021/hash/4cc05b35c2f937c5bd9e7d41d3686fff-Abstract.html.

[12] Zhengzhong Tu, Hossein Talebi, Han Zhang, et al., Maxim: Multi-axis mlp for image processing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5769–5780, http://dx.doi.org/10.1109/CVPR52688.2022.00568.

[13] Noel Codella, Veronica Rotemberg, Philipp Tschandl, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019, arXiv preprint arXiv:1902.03368.

[14] Juan Caicedo, Allen Goodman, Kyle Karhohs, et al., Nucleus segmentation across imaging experiments: the 2018 data science bowl, Nat. Methods 16 (12) (2019) 1247–1253, http://dx.doi.org/10.1038/s41592-019-0612-7.

[15] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, Aly Fahmy, Dataset of breast ultrasound images, in: Data in Brief, Vol. 28, 2020, 104863, http://dx.doi.org/10.1016/j.dib.2019.104863.

[16] Huisi Wu, Shihuai Chen, Guilian Chen, Wei Wang, Baiying Lei, Zhenkun Wen, FAT-Net: Feature adaptive transformers for automated skin lesion segmentation, Med. Image Anal. 76 (2022) 102327, http://dx.doi.org/10.1016/j.media.2021.102327.

[17] Bennett Landman, Zhoubing Xu, J. Igelsias, et al., Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge, Vol. 5, 2015, p. 12, https://www.synapse.org/#!Synapse:syn3193805/wiki/217789.

[18] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440, http://dx.doi.org/10.1038/s41592-019-0612-7.

[19] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, et al., A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986, http://dx.doi.org/10.1109/CVPR52688.2022.01167.

[20] Ze Liu, Yutong Lin, Yue Cao, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022, http://dx.doi.org/10.1109/ICCV48922.2021.00986.

[21] Zhimeng Han, Muwei Jian, Gai-Ge Wang, ConvUNeXt: An efficient convolution neural network for medical image segmentation, Knowl.-Based Syst. 253 (2022) 109512, http://dx.doi.org/10.1016/j.knosys.2022.109512.

[22] Hu Cao, Yueyue Wang, Joy Chen, et al., Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, 2022, pp. 205–218, http://dx.doi.org/10.1007/978-3-031-25066-8_9.

[23] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, et al., Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021, pp. 272–284, http://dx.doi.org/10.1007/978-3-031-08999-2_22.

[24] Hongyu Zhou, Jiansen Guo, Yinghao Zhang, et al., Nnformer: Volumetric medical image segmentation via a 3d transformer, IEEE Trans. Image Process. 32 (2023) 4036–4045, http://dx.doi.org/10.1109/TIP.2023.3293771.

[25] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, et al., UNETR++: delving into efficient and accurate 3D medical image segmentation, 2022, arXiv preprint arXiv:2212.04497.

[26] Jose Jeya Maria Valanarasu, Vishal Patel, Unext: Mlp-based rapid medical image segmentation network, in: International Conference on Medical Image Computing and Computer Assisted Intervention, 2022, pp. 23–33, http://dx.doi.org/10.1007/978-3-031-16443-9_3.

[27] Andrew Howard, Menglong Zhu, Bo Chen, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.

[28] Mark Sandler, Andrew Howard, Menglong Zhu, et al., Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520, http://dx.doi.org/10.1109/CVPR.2018.00474.

[29] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[30] Dan Hendrycks, Kevin Gimpel, Gaussian error linear units (gelus), 2016, arXiv preprint arXiv:1606.08415.

[31] Sergey Ioffe, Christian Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on Machine Learning, Vol. 37, 2015, pp. 448–456, https://proceedings.mlr.press/v162/wang22i.html.

[32] Jimmy Lei Ba, Jamie Ryan Kiros, Geoffrey Hinton, Layer normalization, 2016, arXiv preprint arXiv:1607.06450.

[33] Jianyuan Guo, Yehui Tang, Kai Han, et al., Hire-mlp: Vision mlp via hierarchical rearrangement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 826–836, http://dx.doi.org/10.1109/CVPR52688.2022.00090.

[34] Ziyu Wang, Wenhao Jiang, Yiming M. Zhu, et al., Dynamixer: a vision MLP architecture with dynamic mixing, in: Proceedings of the 39th International Conference on Machine Learning, Vol. 162, 2022, pp. 22691–22701, https://proceedings.mlr.press/v162/wang22i.html.

[35] Chao Ji, Zhaohong Deng, Yan Ding, et al., RMMLP: Rolling MLP and matrix decomposition for skin lesion segmentation, Biomed. Signal Process. Control 84 (2023) 104825, http://dx.doi.org/10.1016/j.bspc.2023.104825.

[36] Huaikun Zhang, Jing Lian, Zetong Yi, et al., HAU-Net: Hybrid CNN-transformer for breast ultrasound image segmentation, Biomed. Signal Process. Control 87 (2024) 105427, http://dx.doi.org/10.1016/j.bspc.2023.105427.

[37] Jingchao Xu, Xin Wang, Wei Wang, Wendi Huang, PHCU-Net: A parallel hierarchical cascade U-Net for skin lesion segmentation, Biomed. Signal Process. Control 86 (2023) 105262, http://dx.doi.org/10.1016/j.bspc.2023.105427.

[38] Cliff Rosendahl, Harald Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (1) (2018) 1–9, http://dx.doi.org/10.1038/sdata.2018.161.

[39] Qing Xu, Zhicheng Ma, Na H.E., Wenting Duan, DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation, Comput. Biol. Med. 154 (2023) 106626, http://dx.doi.org/10.1016/j.compbiomed.2023.106626.

[40] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, Jianming Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 3–11, http://dx.doi.org/10.1007/978-3-030-00889-5_1.

[41] Debesh Jha, Pia Smedsrud, Michael Riegler, et al., Resunet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia, 2019, pp. 225–2255, http://dx.doi.org/10.1109/ISM46123.2019.00049.

[42] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, et al., Attention u-net: Learning where to look for the pancreas, 2018, arXiv preprint arXiv:1804.03999.

[43] Yongheng Sun, Duwei Dai, Qianni Zhang, et al., MSCA-Net: Multi-scale contextual attention network for skin lesion segmentation, Pattern Recognit. 139 (2023) 109524, http://dx.doi.org/10.1016/j.patcog.2023.109524.

[44] Wenyu Zhang, Fuxiang Lu, Wei Zhao, et al., ACCPG-Net: A skin lesion segmentation network with adaptive channel-context-aware pyramid attention and global feature fusion, Comput. Biol. Med. 154 (2023) 106580, http://dx.doi.org/10.1016/j.compbiomed.2023.106580.

[45] Bingzhi Chen, Yishu Liu, Zheng Zhang, et al., Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation, IEEE Trans. Emerg. Top. Comput. Intell. 8 (1) (2024) 55–68, http://dx.doi.org/10.1109/TETCI.2023.3309626.

[46] Reza Azad, René Arimond, Ehsan Khodapanah Aghdam, et al., Dae-former: Dual attention-guided efficient transformer for medical image segmentation, in: International Workshop on PRedictive Intelligence in MEdicine, Vol. 14277, Springer Nature Switzerland, Cham, 2023, pp. 83–95, http://dx.doi.org/10.1007/978-3-031-46005-0_8.

[47] Yading Yuan, Ming Chao, Yeh-Chi Lo, Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance, IEEE Trans. Med. Imaging 36 (9) (2017) 1876–1886, http://dx.doi.org/10.1109/TMI.2017.2695227.

[48] William Crum, Oscar Camara, Derek Hill, Generalized overlap measures for evaluation and validation in medical image analysis, IEEE Trans. Med. Imaging 25 (11) (2006) 1451–1461, http://dx.doi.org/10.1109/TMI.2006.880587.

[49] Nikhil Kumar Tomar, Debesh Jha, Michael Riegler, et al., FANet: A feedback attention network for improved biomedical image segmentation, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–14, http://dx.doi.org/10.1109/TNNLS.2022.3159394.

[50] Nikhil Kumar Tomar, Debesh Jha, Sharib Ali, et al., DDANet: Dual decoder attention network for automatic polyp segmentation, Pattern Recognit. (2021) 307–314, http://dx.doi.org/10.1007/978-3-030-68793-9_23.

[51] Debesh Jha, Michael A. Riegler, Dag Johansen, et al., Doubleu-net: A deep convolutional neural network for medical image segmentation, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems, 2020, pp. 558–564, http://dx.doi.org/10.1109/CBMS49503.2020.00111.

[52] Kim Taehun, Hyemin Lee, Daijin Kim, Uacanet: Uncertainty augmented context attention for polyp segmentation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2167–2175, http://dx.doi.org/10.1145/3474085.3475375.

[53] Abhishek Srivastava, Debesh Jha, Sukalpa Chanda, et al., MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation, IEEE J. Biomed. Health Inf. 26 (5) (2021) 2252–2263, http://dx.doi.org/10.1109/JBHI.2021.3138024.

[54] Michal Byra, Piotr Jarosik, Aleksandra Szubert, et al., Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network, Biomed. Signal Process. Control 61 (2020) 102027, http://dx.doi.org/10.1016/j.bspc.2020.102027.

[55] Gongping Chen, Lei Li, Yu Dai, et al., AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images, IEEE Trans. Med. Imaging 42 (5) (2022) 1289–1300, http://dx.doi.org/10.1109/TMI.2022.3226268.

[56] Chenyu You, Ruihan Zhao, Fenglin Liu, et al., Class-aware adversarial transformers for medical image segmentation, Adv. Neural Inf. Process. Syst. 35 (2022) 29582–29596, https://proceedings.neurips.cc/paper_files/paper/2022/hash/be99227ef4a4de84bb45d7dc7b53f808-Abstract-Conference.html.

[57] Hengshuang Zhao, Li Jiang, Jiaya Jia, et al., Point transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16259–16268, http://dx.doi.org/10.1109/ICCV48922.2021.01595.

[58] Xuyang Bai, Zixin Luo, Lei Zhou, et al., Pointdsc: Robust point cloud registration using deep spatial consistency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15859–15869, http://dx.doi.org/10.1109/CVPR46437.2021.01560.