PL-Net: progressive learning network for medical image segmentation

Kunpeng Mao[1†], Ruoyu Li[2†], Junlong Cheng[2], Danmei Huang[1], Zhiping Song[3]* and ZeKui Liu[3]*

[1]Chongqing City Management College, Chongqing, China, [2]College of Computer Science, Sichuan University, Chengdu, China, [3]Chongqing University of Engineering, Chongqing, China

In recent years, deep convolutional neural network-based segmentation methods have achieved state-of-the-art performance for many medical analysis tasks. However, most of these approaches rely on optimizing the U-Net structure or adding new functional modules, which overlooks the complementation and fusion of coarse-grained and fine-grained semantic information. To address these issues, we propose a 2D medical image segmentation framework called Progressive Learning Network (PL-Net), which comprises Internal Progressive Learning (IPL) and External Progressive Learning (EPL). PL-Net offers the following advantages: 1) IPL divides feature extraction into two steps, allowing for the mixing of different size receptive fields and capturing semantic information from coarse to fine granularity without introducing additional parameters; 2) EPL divides the training process into two stages to optimize parameters and facilitate the fusion of coarse-grained information in the first stage and fine-grained information in the second stage. We conducted comprehensive evaluations of our proposed method on five medical image segmentation datasets, and the experimental results demonstrate that PL-Net achieves competitive segmentation performance. It is worth noting that PL-Net does not introduce any additional learnable parameters compared to other U-Net variants.

1 Introduction

Medical image segmentation is a technique used to extract regions of interest for quantitative and qualitative analysis. For example, it can be used for cell segmentation in electron microscopy (EM) recordings (Ronneberger et al., 2015), melanoma segmentation in dermoscopy images (Berseth, 2017; Cheng et al., 2020), thyroid nodule segmentation in ultrasound images, and heart segmentation in MRI images (Bernard et al., 2018). Traditionally, medical image segmentation methods relied on manually designed features to generate segmentation results (Xu et al., 2007; Tong et al., 2015). However, this approach requires distinct feature designs for various applications. Furthermore, the large variety of medical image modalities makes it difficult or impossible to transfer a specific type of feature design method to different image types. Therefore, the development of a universal feature extraction technique is crucial in the field of medical image analysis.

The emergence of deep learning technology has revolutionized medical image segmentation by overcoming the limitations of traditional manual feature extraction methods. Convolutional neural networks (CNN) based automatic feature learning algorithms, such as the fully convolutional network (FCN) proposed by Shelhamer et al. (Long et al., 2015) and the U-Net framework for biomedical image segmentation proposed by Ronneberger et al., 2015, have shown promising results. The FCN model structure is designed to be end-to-end, which eliminates the need for manual feature extraction and image post-processing steps. On the other hand, the U-Net framework's encoder-decoder-skip connection network structure has shown good results on medical image segmentation datasets with small amounts of data.

To further enhance the adaptability of U-Net for different medical image segmentation tasks, researchers have continuously explored and innovated, proposing numerous variant models of U-Net. These variant models aim to achieve better performance in medical image segmentation by adding new functional modules or optimizing the network structure. For instance, Vanilla U-Net introduces channel/spatial attention mechanisms or self-attention mechanisms to capture crucial information in medical images, significantly improving its performance in various segmentation tasks. Additionally, researchers have optimized the encoder-decoder structure of Vanilla U-Net or adjusted the skip connections to generate more refined and abundant feature representations.

However, the introduction of these methods has also brought new challenges. Although the addition of new parameters and functional modules enhances model performance, it also increases model complexity and the risk of overfitting. More importantly, these methods often overlook the complementarity and fusion of coarse-grained and fine-grained semantic information. Most existing semantic segmentation methods assume that the entire segmentation process can be completed through a single feedforward process, resulting in homogeneous feature representations that struggle to excel in extracting fine-grained feature representations. Therefore, for medical environments with limited computational resources, it is highly beneficial to ensure model simplicity while fully integrating and utilizing semantic information at different scales while maintaining a small number of parameters. Such a design can not only enhance the generalization and robustness of the model but also ensure its efficiency and practicality in real-world applications.

In this paper, we propose a new medical image segmentation method called progressive learning networks (PL-Net). PL-Net divide the feature learning process within the U-Net architecture into two distinct depth "steps" to achieve the combination of different receptive field sizes, enabling the network to learn semantic information at varying granularities. The entire segmentation process is performed through two feedforward processes (referred to as "stages"). At the end of each stage, the features obtained from that stage are transferred to the next stage for fusion. This transfer operation allows the model to leverage the knowledge learned in the previous training stage to extract finer-grained information, thereby refining the coarse segmentation output. Unlike previous works, our proposed method does not add any additional parameters or functional modules to the U-Net. Instead, our method fully explores the complementary

relationships between features through a progressive learning strategy. The main contributions can be summarized as follows:

1) We propose a progressive learning network (PL-Net) designed specifically for medical image segmentation tasks. Through its unique design, this network deeply explores the potential of feature complementarity and fusion in medical image segmentation. By adjusting the scale of output channels, we designed both a standard PL-Net (15.03 M) and a smaller version, PL-Net† (Ocs = 0.5, 3.77 M), to accommodate medical scenarios with different computational resources.

2) We introduce internal progressive learning (IPL) and external progressive learning (EPL) strategies. The IPL strategy effectively captures different receptive field sizes, thereby learning and integrating multi-granularity semantic information. The EPL strategy allows the model to extract finer information based on the knowledge from the previous stage, thus optimizing the segmentation results.

3) We applied the proposed method to tasks such as skin lesion segmentation and cell nucleus segmentation. Experimental results indicate that PL-Net outperforms other state-of-the-art methods such as U-NeXt and BiO-Net. Moreover, despite its smaller parameter size, the smaller version of PL-Net† still demonstrates superior segmentation performance.

2 Releat work

Currently, most semantic segmentation methods assume that the entire segmentation process can be executed through a single feedforward pass of the input image, which often overlooks global information. To address this, researchers have added new functional modules or optimized the U-Net structure to achieve performance improvements. These methods can be classified into: 1) U-Net variants focused on functional optimization; 2) U-Net variants focused on structural optimization.

U-Net variants focused on functional optimization. Due to the large number of irrelevant features in medical images, it is crucial to focus on target features and suppress irrelevant features during the segmentation process. Recent works have extended U-Net by adding different novel functional modules, demonstrating its potential in various visual analysis tasks. Squeeze-and-Excitation (SE) (Hu et al., 2018) has facilitated the development of U-Net by automatically learning the importance of each feature channel through an attention mechanism. Additionally, ScSE (Roy et al., 2018) and FCANet (Cheng et al., 2020) integrate concurrent spatial and channel attention modules into U-Net to improve segmentation performance. Oktay et al., 2018 proposed an attention gate for medical imaging to focus on target structures of different shapes and sizes and suppress irrelevant areas of the input image. In addition to plug-and-play attention modules, researchers have designed specific functional modules for different medical image segmentation tasks. For example, Zhou et al., 2019 proposed a contour-aware information aggregation network with a multi-level information aggregation module between two task-specific decoders. The SAUNet (Sun et al., 2020) uses both a secondary shape stream and a regular texture stream in parallel to capture rich shape-related information, enabling multi-level interpretation of the external

network and reducing the need for additional computations. The CE-Net (Gu et al., 2019) uses a dense atrous convolution (DAC) block to extract a rich feature representation and residual multi-kernel pooling (RMP) operation to further encode the multi-scale context features extracted from the DAC block without additional learning weights.

The emergence of the Vision Transformer (ViT) (Dosovitskiy et al., 2020) has had a significant impact on the progress of medical image analysis. Compared to CNN methods, ViT has less inductive bias. The U-Transformer (Petit et al., 2021) takes inspiration from ViT and incorporates multi-head self-attention modules into U-Net, which helps to obtain global contextual information. The UNeXt (Valanarasu and Patel, 2022) is the first fast medical image segmentation network that uses both convolution and MLP. It reduces the number of parameters and computational complexity by using tokenized MLP. In contrast to the aforementioned U-Net variants, our work explores the effectiveness of progressive learning techniques in capturing both coarse-grained and fine-grained semantic information. The PL-Net enhances the performance of different stage U-Nets by reusing learned features.

U-Net variants focused on structural optimization. Unlike U-Net variants focused on functional optimization, optimizing its structure allows it to extract feature information at different levels, which is feasible and effective for many computer vision problems. One of the simplest and most effective ways to optimize the encoder-decoder structure is to replace the basic building blocks with more advanced ones, such as (Jégou et al., 2017; Diakogiannis et al., 2020; Hasan et al., 2020), which benefit from residual or dense connections in deeper network structures. In addition to replacing the building blocks, performance can also be improved for different tasks by increasing the number of U-shaped network structures, as demonstrated in (Jégou et al., 2017; Isensee et al., 2021). One of the most famous networks in this category is nnU-Net (Isensee et al., 2021), which proposes three networks based on the original U-Net structure: 2D U-Net, 3D U-Net, and U-Net cascade. The first stage performs coarse segmentation of downsampled low-resolution images, and the second stage combines the results of the first stage for fine-tuning. ResGANet (Cheng et al., 2022) achieved segmentation performance improvement by replacing the encoder in U-Net with a lightweight and efficient backbone. TransUNet (Chen J. et al., 2021) and FATNet (Wu et al., 2022) replaced the encoder structure of U-Net with CNN and Transformer branches in a parallel or serial manner.

Skip connections are considered a key component of U-Net's success. U-Net++ (Zhou et al., 2018) has redesigned skip connections through a series of nested and dense connections, reducing the semantic gap between the subnet feature map encoders and decoders. R2U-Net (Alom et al., 2019) effectively increases the network depth by utilizing residual networks and RCNN and obtaining more expressive features through feature summation with different time steps. Xiang et al. designed BiO-Net (Xiang et al., 2020) with backward skip connections based on R2UNet, which can reuse the features of each decoding level to achieve more intermediate information aggregation. The emergence of BiO-Net allows building blocks to be reused by U-Net in a circular manner without introducing any additional parameters.

3 Progressive learning network

We now describe PL-Net, a progressive learning framework for medical image segmentation. As is shown in Figure 1, PL-Net is a multi-level U-Net network architecture that does not rely on additional functional modules but has paired bidirectional connections. The core of our framework is to enhance the feature representation required for image segmentation through two progressive learning approaches (internal and external) and to fuse coarse-grained as well as fine-grained semantic information.

Two U-Nets with different depths form the "stages" of external progressive learning. In each stage, as the "step" of internal progressive learning increases, the shallow network is expanded to a deeper network, learning stable multi-granularity information from it. In brief, the number of parameters is not increased through internal progressive learning, but it can learn feature maps with different sizes of receptive fields. External progressive learning is defined as the coarse segmentation stage (Stage 1) and the fine segmentation stage (Stage 2). The input image will be examined at multiple scales to achieve the fusion of coarse-grained and fine-grained information.

3.1 Internal progressive learning

Bidirectional skip connections are used in internal progressive learning to reuse building blocks. In order to enable the network at each stage to learn distinctive feature representations, we use two "steps" to gradually mine the features from shallow to deep.

Forward Skip Connections (FSC) are used to assist up-sampling learning, restore the contour of the segmentation target, and retain the low-level visual features of encoding. The feature f_s^{FSC} after FSC can be expressed as:

$$f_s^{FSC} = [Conv_s(x), \hat{x}] \tag{1}$$

Backward Skip Connections (BSC) are used for flexible aggregation of low-level visual features and high-level semantic features. In order to realize the complementation and fusion of semantic information at different stages, multi-granularity information of different "steps" and "stages" is combined in feature f^{BSC}:

$$f^{BSC} = \begin{cases} [x, x] & s = 1, stage1 \\ [Conv_s(f_s^{FSC}), x] & s = 1, stage2 \\ [Conv_s(\hat{x}), x] & s = 2, stage1, 2 \end{cases} \tag{2}$$

Among them, $[\cdot]$ is the concatenation layer, $Conv_s$ means that the convolution operation of the sth "steps" ($s \in \{1, 2\}$) is applied to the input feature map, x and $\hat{x}$ are feature maps of the same size in the down-sampling and up-sampling path respectively.

It is worth noting that the reasoning path of internal progressive learning can be extended to multiple recursions to obtain instant performance gains. More importantly, a larger receptive field will be got in each output of this learning strategy than the previous "steps." We use K_i to represent the ith complete encoding-decoding process, and x_{out}^i is used to represent the output. Therefore, x_{out}^i can be written as:
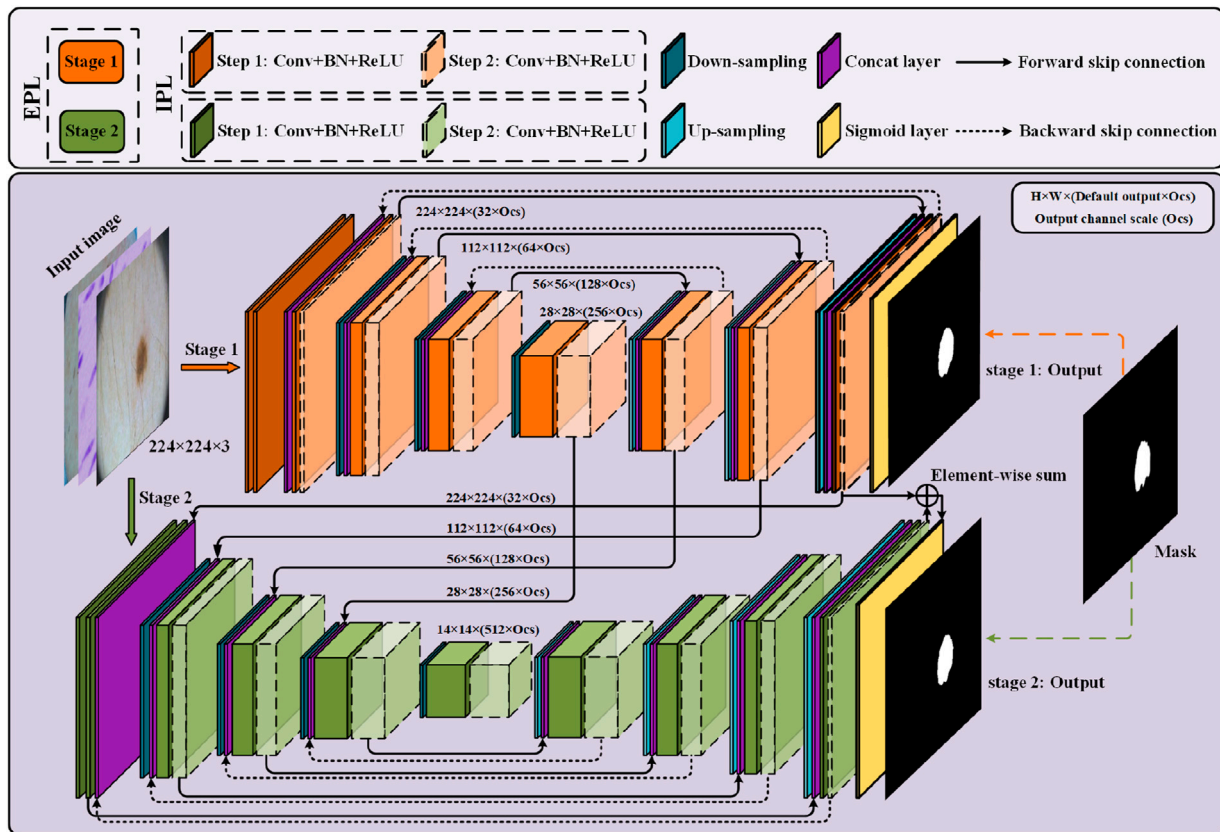
FIGURE 1
Overview of the progressive learning network (PL-Net). PL-Net consists of two parts: internal progressive learning (IPL) and external progressive learning (EPL).

$$x_{out}^{i} = \begin{cases} x_{in} & i = 0 \\ K_i(x_{in}) & i = 1 \\ K_i(x_{out}^{i-1}) & i \geq 2 \end{cases} \quad (3)$$

In this study, we define $i = 2$, and through such a setting the parameters equivalent to BiO-Net can be maintained. In future research, the setting of $i > 2$ can be used to further improve the segmentation accuracy, but the exploration of the optimal hyperparameter setting is beyond the scope of this paper.

3.2 External progressive learning

The external progressive learning strategy first trains the low stage (stage 1), and then gradually trains toward the high stage (stage 2). Since "stage1″ is relatively shallow in depth and limited by the perceptual field and performance ability, it will focus on local information extraction, while "stage 2″ incorporates the local texture information learned from "stage 1." Compared with directly training the entire network in series, in the model, it is allowed by this incremental nature to pay attention to global information as the features gradually enter a higher stage.

For each stage of training, we calculate the loss based on the Dice coefficient ($\mathcal{L}_{Dice}$) (Eelbode et al., 2020) between the ground truth (y_{true}) and the predicted probability (y_{pred}^n) distribution of different stages:

$$\mathcal{L}_{Dice}\left(y_{pred}^n, y_{true}\right) = 1 - \frac{2 \times |y_{pred}^n \cap y_{true}|}{|y_{pred}^n| + |y_{true}|} \quad (4)$$

Here $|\cdot|$ is an operator through which the number of pixels is found in the qualified area. In each iteration, the input data will be used in each learning stage (where $n \in \{1, 2\}$). What needs to be clear is that when the latter stage is predicted, all the parameters of the previous stage are optimized and updated, which helps each stage in the model to work together.

Since the low stage is mainly to assist the feature expression and knowledge reasoning of the high-stage network, we can delete the low-stage prediction layer (Sigmoid layer) when predicting, thereby reducing the reasoning time. In addition, the predictions at different stages are unique, but they can form complementary information among different granularities. When we combine all outputs with an equal weight, it will result in a better performance. In other words, the final output of the model is determined by all stages:

$$y = \frac{1}{1 + e^{-\sum_{n=1}^{N-1} y^n}} \quad (5)$$

3.3 PL-Net architecture

Our framework has a trade-off between performance and parameters. Like U-Net, the down-sampling and up-sampling

TABLE 1 Detailed configuration of U-Net, BiO-Net, and our PL-Net architecture. We use " [kernel, kernel, channel]" to represent the convolution configuration.

Input	Encoder			Output	Decoder		
	U-Net	BiO-Net	PL-Net		U-Net	BiO-Net	PL-Net
224^2	$[3, 3, 64]\times 2$	$[3, 3, 32]\times 2$	$\begin{Bmatrix}[3,3,32],\ \text{step1}\\ [3,3,32],\ \text{step2}\\ \text{stage1, stage2}\end{Bmatrix}$	7^2	—	$[3, 3, 256]\times 2$	—
112^2	$[3, 3, 128]\times 2$	$[3, 3, 32]\times 2$	$\begin{Bmatrix}[3,3,64],\ \text{step1}\\ [3,3,64],\ \text{step2}\\ \text{stage1, stage2}\end{Bmatrix}$	28^2	$[3, 3, 512]\times 2$	$[3, 3, 128]\times 2$	$\begin{Bmatrix}[3,3,256],\ \text{step1}\\ [3,3,256],\ \text{step2}\\ \text{stage2}\end{Bmatrix}$
56^2	$[3, 3, 256]\times 2$	$[3, 3, 64]\times 2$	$\begin{Bmatrix}[3,3,128],\ \text{step1}\\ [3,3,128],\ \text{step2}\\ \text{stage1, stage2}\end{Bmatrix}$	56^2	$[3, 3, 256]\times 2$	$[3, 3, 64]\times 2$	$\begin{Bmatrix}[3,3,128],\ \text{step1}\\ [3,3,128],\ \text{step2}\\ \text{stage1, stage2}\end{Bmatrix}$
28^2	$[3, 3, 512]\times 2$	$[3, 3, 128]\times 2$	$\begin{Bmatrix}[3,3,256],\ \text{step1}\\ [3,3,256],\ \text{step2}\\ \text{stage1, stage2}\end{Bmatrix}$	112^2	$[3, 3, 128]\times 2$	$[3, 3, 32]\times 2$	$\begin{Bmatrix}[3,3,64],\ \text{step1}\\ [3,3,64],\ \text{step2}\\ \text{stage1, stage2}\end{Bmatrix}$
14^2	$[3, 3, 1024]\times 2$	$[3, 3, 256]\times 2$	$\begin{Bmatrix}[3,3,512],\ \text{step1}\\ [3,3,512],\ \text{step2}\\ \text{stage2}\end{Bmatrix}$	224^2	$[3, 3, 64]\times 2$	$[3, 3, 32]\times 2$	$\begin{Bmatrix}[3,3,32],\ \text{step1}\\ [3,3,32],\ \text{step2}\\ \text{stage1, stage2}\end{Bmatrix}$
7^2	—	$[3, 3, 512]\times 2$	—	224^2	$[1, 1, 1]$, *Sigmoid*		
Parameters					25.59 M	14.30 M	15.03 M
Model size					118 MB	57.7 MB	60.7 MB

stages of PL-Net only use standard convolutional layers, batch normalization layers and ReLU layers without introducing any additional functional modules. Table 1 is the detailed configuration of U-Net, BiO-Net and our PL-Net.

As shown in Table 1, BiO-Net has a maximum coding depth of 4, using BSC from the deepest coding level, and inputting the decoded features in each iteration as a whole into the last-stage block. BSC is also used in PL-Net. Unlike the previous methods, the convolutional layer is allowed to be used in the model to mine features from coarse-grained to fine-grained ones in a progressive manner. It should be noted that a smaller version of PL-Net† can be obtained only by adjusting the Ocs, whose depth and connection method will not change.

4 Experiments

4.1 Datasets

ISIC 2017 (Berseth, 2017) is a dataset consisting of 2000 training images, 150 validation images, and 600 test images. The images in the original dataset provided by ISIC have different resolutions. To address this, we first use the gray world color consistency algorithm to normalize the colors of the images and then adjust the size of all images to 2242 pixels. All experimental results reported in this paper for ISIC 2017 are from the official test set results.

PH2 (Mendonça et al., 2013) is a dataset containing 200 dermoscopic images, with a fixed size of 768×560 pixels. The dataset contains 80% benign mole cases and 20% melanoma cases, with ground truth annotated by dermatologists. Due to the small scale of the dataset, we use the preprocessing method of the ISIC 2017 dataset and the trained model to directly predict all images in the dataset to evaluate the performance of different models.

Kaggle 2018 Data Science Bowl (referred to as Nuclei) (Caicedo et al., 2019) is a dataset provided by the Booz Allen Foundation, containing 670 cell feature maps with ground truth for each image. To prepare the dataset for training and testing, we adjust all images and corresponding ground truth to 2242 pixels and use 80% of the images for training and the remaining 20% for testing.

The TN-SCUI (Pedraza et al., 2015) dataset is a collection of 3644 nodular thyroid images, each annotated by experienced physicians. The dataset was originally part of the TN-SCUI 2020 challenge and was processed to remove personal labels to protect patient privacy. In this study, we randomly divided the dataset into a training set (60%), validation set (20%), and test set (20%). To ensure consistency, we uniformly adjusted the resolution of all images to 2242 pixels.

ACDC (Bernard et al., 2018) is a dataset that includes cardiac MRI images of 150 patients, from which we collected 1489 slices for 3D images. For training and testing purposes, we used 951 and 538 slices, respectively. Notably, in contrast to the four other datasets mentioned earlier, ACDC comprises three different categories: left ventricle, right ventricle, and myocardium. Hence, we employed this dataset to investigate how various models perform on multi-class segmentation. Figure 2 displays sample images from these datasets and their corresponding ground truth.

4.2 Implementation details

We conducted all experiments on Tesla V100 GPUs using Keras and expanded the training data for all datasets by applying random rotations (±25°), random horizontal and vertical shifts (15%), and random flips (horizontal and vertical). For all models, we trained for more than 200 epochs with a batch size of 16, a fixed learning rate of 1e-4, and an Adam optimizer with a momentum of 0.9 to minimize
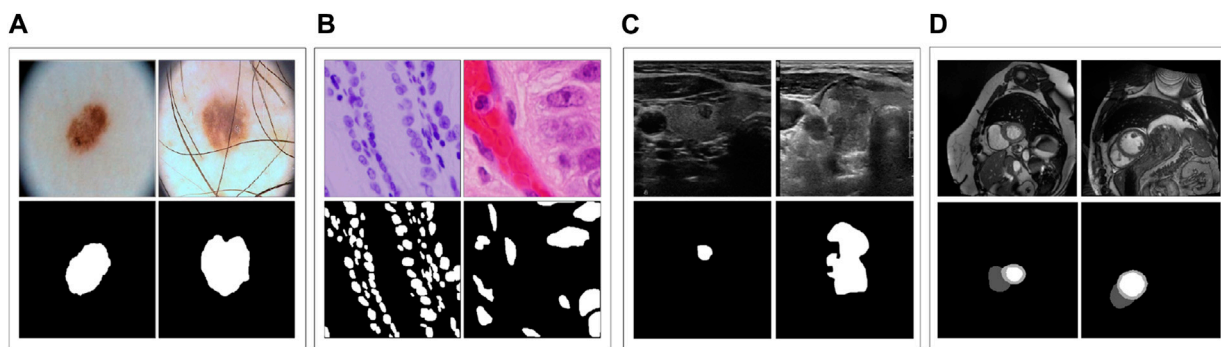
FIGURE 2
(A–D) represent samples from the five datasets.

TABLE 2 Ablative results. IoU (Dice), number of parameters, and model size are reported.

Dataset	EPL	Without IPL		With IPL	
		n = 1	*n* = 2	*n* = 2	*n* = 3
ISIC 2017	×	76.05 (84.43)	77.09 (85.23)	77.15 (85.37)	77.07 (85.44)
	✓	76.69 (84.94)	77.04 (85.27)	**77.92 (85.94)**	77.49 (85.56)
Nuclei	×	85.54 (91.89)	85.13 (91.53)	85.93 (**92.14**)	84.78 (91.28)
	✓	85.60 (91.84)	85.80 (92.00)	**86.14** (92.13)	85.37 (91.70)
PH2	×	83.90 (90.74)	85.88 (91.61)	86.69 (92.47)	86.48 (92.45)
	✓	84.97 (91.00)	86.63 (92.44)	**87.27 (92.86)**	87.03 (92.77)
TN-SCUI	×	72.32 (81.38)	73.72 (82.61)	75.95 (85.36)	76.67 (84.60)
	✓	74.20 (83.33)	75.63 (84.33)	76.66 (85.10)	**77.05 (85.55)**
ACDC	×	74.44 (81.56)	74.66 (82.03)	77.78 (83.84)	77.49 (83.60)
	✓	75.30 (82.19)	76.80 (83.42)	**78.06 (84.36)**	77.96 (83.91)
Parameters	—	**10.33M**	15.03 M	15.03 M	19.73 M
Model size	—	**41.60 MB**	60.70 MB	60.70 MB	79.70 MB

The bold values indicates the best performance.

Dice loss. We used an early stop mechanism and stopped training when the validation loss reached a stable level with no significant change for 20 epochs. Unless explicitly specified, PL-Net had two "steps" and "stages," and BSC was established at each stage of the network. When testing, all prediction layers are deleted before the last "stage," and other configurations remain unchanged.

4.3 Ablation study

To understand the effectiveness of IPL and EPL strategies, we conducted ablation studies. When there is no IPL strategy, features are extracted by naturally stacking benchmark blocks, and we conducted experiments on stacking 1-layer and 2-layer benchmark blocks, respectively. Adopting an IPL strategy means that the encoder-decoder must be iterated for n times in different stages, and we set *n* = 2 and *n* = 3. When external progressive learning is not performed, different "stages" are connected in series

through PL-Net to transfer the feature information learned in each stage. Only the parameters in the last stage are optimized, and the segmentation results are output through the model. That is to say, the feature information obtained in the current "stage" of training is transferred to the next training "stage" and fused through the EPL method, allowing fine-grained information to be mined through the model based on learning in the previous training "stage".

Table 2 presents our IoU (Dice) scores without/with a progressive learning strategy on five different medical image segmentation datasets. We provide the parameters and model sizes for different scenarios to comprehensively analyze segmentation performance. In most cases, the best segmentation performance is achieved through PL-Net when both internal (*n* = 2) and external progressive learning strategies are used simultaneously. Compared to the model with the same parameter settings without IPL, the segmentation performance is significantly improved. These results demonstrate the effectiveness of EPL and IPL. Moreover, we observed that the progressive learning strategy has a significant
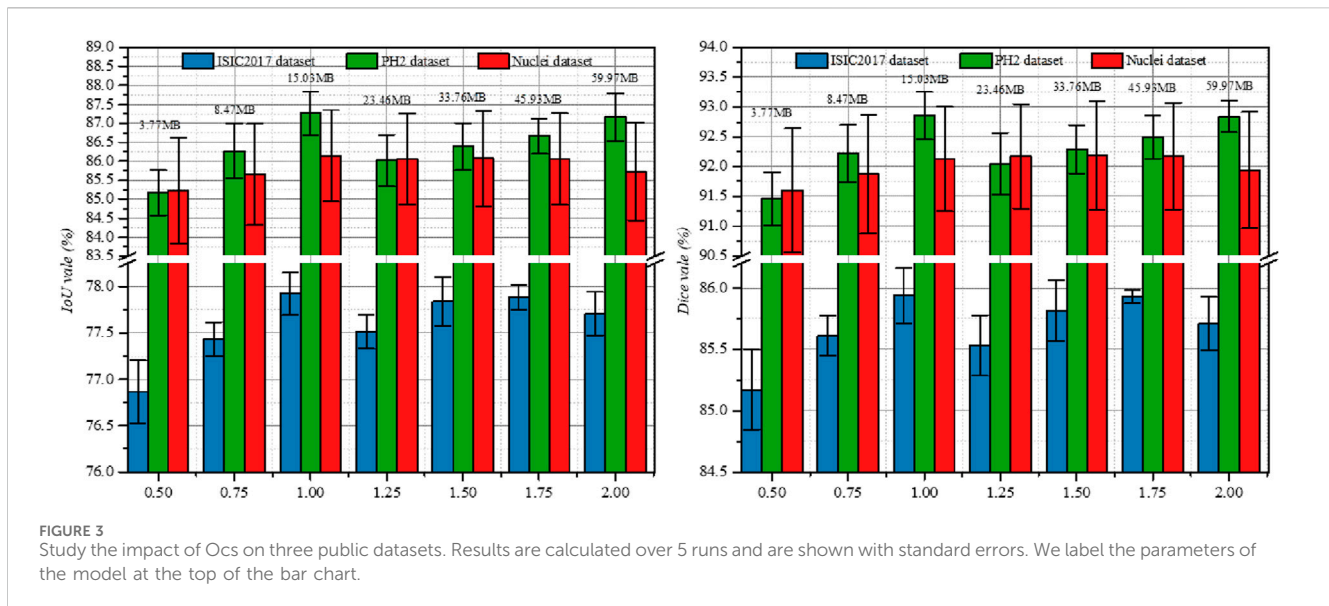
FIGURE 3
Study the impact of Ocs on three public datasets. Results are calculated over 5 runs and are shown with standard errors. We label the parameters of the model at the top of the bar chart.

impact on datasets with complex boundaries or multi-category datasets. On the TN-SCUI dataset, for instance, the IoU improvement is as high as 2.94% with the same parameter setting (n = 2). To balance factors such as performance and parameters, we used the setting of n = 2 in the following experiments. However, we believe that the setting of n = 3 may be more effective as the size of the dataset increases.

In addition to the above ablation studies, we also investigated the impact of the output channel scale (Ocs) on the segmentation performance of different datasets. Figure 3 shows the experimental results on three datasets, where we set $Ocs \in$ [0.5, 2.0] and take values at an interval of 0.25. Note that Ocs = 0.5 represents a smaller version of PL-Net†.We found that when Ocs = 1.0, the best segmentation result can be obtained, and the parameter amount (15.03 MB) is well-balanced. When Ocs > 1.0, the segmentation performance improves as the number of channels increases, but it does not exceed that of the standard PL-Net. We attribute this to the limitation of the data size and the complexity of the segmentation content. While PL-Net† has slightly lower segmentation performance than other networks, it has very few parameters. Thus, it is recommended for use on small datasets. Additionally, it can be configured to run on servers or mobile devices with lower hardware requirements.

4.4 Comparison with state-of-the-arts

4.4.1 Quantitative comparison

For the ISIC 2017 and PH2 datasets, we compared our PL-Net to the baseline U-Net and other state-of-the-art methods (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Al-Masni et al., 2018; Oktay et al., 2018; Zhou et al., 2018; Alom et al., 2019; Kaul et al., 2019; Cheng et al., 2020; Hasan et al., 2020; Jha et al., 2020; Lei et al., 2020; Xiang et al., 2020; Cao et al., 2021; Isensee et al., 2021; Cheng et al., 2022; Valanarasu and Patel, 2022; Wu et al., 2022). The functional optimization-oriented variants of U-Net include (Badrinarayanan et al., 2017; Al-Masni et al., 2018; Oktay et al., 2018; Alom et al.,

2019; Kaul et al., 2019; Cheng et al., 2020; Valanarasu and Patel, 2022) while the structural optimization-oriented variants of U-Net include (Zhou et al., 2018; Alom et al., 2019; Hasan et al., 2020; Jha et al., 2020; Lei et al., 2020; Xiang et al., 2020; Cao et al., 2021; Isensee et al., 2021; Cheng et al., 2022; Wu et al., 2022). To ensure fairness, we either used the experimental results provided by the authors on the same test set or ran their models published in the same environment.

Table 3 presents the accuracy (Acc), intersection over union (IoU), Dice coefficient (Dice), sensitivity (Sens), and specificity (Spec) scores of different segmentation methods on the ISIC2017 and PH2 datasets. Our PL-Net outperforms other methods in terms of both IoU and Dice metrics on the ISIC2017 dataset. Specifically, the IoU and Dice scores of PL-Net are 0.6% and 0.3% higher than those of BiO-Net (t = 3, INT), respectively. The smaller sized PL-Net† (3.77 M) achieves the same Dice score as BiO-Net (t = 3) (14.30 M). Although nnU-Net (Petit et al., 2021) achieves the best sensitivity on the ISIC2017 test set, its model size is 3.76 times larger than that of the standard PL-Net. The PH2 dataset also involves the dermoscopic image segmentation task. While the number of parameters in UNeXt-L (Dosovitskiy et al., 2020) is similar to that of our smaller version of PL-Net†, UNeXt-L completes the entire segmentation process through a single feed-forward pass of the input image, resulting in low parameter utilization and insufficient learning. When compared with other state-of-the-art methods, PL-Net demonstrates superior performance on the PH2 dataset. Furthermore, PL-Net† hasmuch fewer parameters than other methods, yet it still achieves competitive segmentation performance.

Nuclei dataset. The datasets used for nucleus segmentation have non-uniform feature distributions, and the shapes of positive and negative samples vary greatly. Table 4 presents the quantitative comparison results of our method against 14 other methods. Compared to the state-of-the-art TransAttUnet-R (Chen B. et al., 2021), our PL-Net achieves better overall segmentation performance, with improvements ranging from 0.27% to 1.16% for different evaluation metrics. The segmentation performance

TABLE 3 Performance comparison with SOTA methods on ISIC 2017 and PH2 datasets. Red, Green, and Blue indicate the best, second best and third best performance.

Network	ISIC 2017 dataset					PH2 dataset					#Params	Model size
	Acc	IoU	Dice	Sens	Spec	Acc	IoU	Dice	Sens	Spec		
FrCN Al-Masni et al. (2018)	0.940	0.771	0.871	0.854	0.967	0.951	0.848	0.918	0.937	0.957	—	—
FocusNet Tong et al. (2015)	0.921	0.756	0.832	0.767	0.990	—	—	—	—	—	—	—
SegNet Kaul et al. (2019)	0.918	0.696	0.821	0.801	0.954	0.934	0.808	0.894	0.865	0.966	28.09M	112 MB
DSNet Gu et al. (2019)	—	0.775	—	0.875	0.967	—	0.870	—	0.929	0.969	10.00 M	—
DAGAN Valanarasu and Patel (2022)	0.935	0.771	0.859	0.835	0.976	—	—	—	—	—	—	—
FATNet Wu et al. (2022)	0.933	0.765	0.850	0.839	0.973	—	—	—	—	—	27.43 M	109 MB
ResGANet Cheng et al. (2022)	0.936	0.764	0.862	0.842	0.950	—	—	—	—	—	39.21M	—
U-Net Ronneberger et al. (2015)	0.926	0.736	0.825	0.828	0.964	0.943	0.851	0.915	0.946	0.957	29.59 M	118 MB
U-Net++ Berseth (2017)	0.929	0.753	0.840	0.848	0.965	0.948	0.853	0.917	0.973	0.937	34.48 M	138 MB
Double U-Net Isensee et al. (2021)	0.936	0.765	0.847	0.830	0.970	0.942	0.860	0.915	0.934	0.953	27.94 M	112 MB
FCANet Hu et al. (2018)	0.935	0.776	0.856	0.869	0.962	0.952	0.868	0.926	0.968	0.926	59.97 M	241 MB
Att U-Net Cheng et al. (2020)	0.939	0.757	0.840	0.859	0.957	0.951	0.868	0.926	0.953	0.955	30.42 M	121 MB
R2U-Net Cheng et al. (2022)	0.938	0.776	0.858	0.859	0.969	0.952	0.871	0.927	0.954	0.960	91.61 M	366 MB
Att R2U-Net Cheng et al. (2022)	0.939	0.775	0.857	0.857	0.961	0.954	0.873	0.928	0.949	0.962	92.11 M	368 MB
BiO-Net (t = 3) Zhou et al. (2018)	0.937	0.772	0.852	0.845	0.973	0.944	0.851	0.915	0.963	0.931	14.30 M	57.7 MB
BiO-Net (t = 3, INT) Zhou et al. (2018)	0.934	0.754	0.840	0.821	0.976	0.945	0.851	0.909	0.968	0.944	3.99 M	15.2 MB
UNeXt-L Valanarasu and Patel (2022)	0.935	0.773	0.856	0.864	0.966	0.949	0.859	0.919	0.941	0.955	14.30 M	57.7 MB
PL-Net† (Our)	0.932	0.769	0.852	0.871	0.953	0.945	0.852	0.915	0.955	0.957	3.77 M	15.6 MB
PL-Net (Our)	0.940	0.779	0.859	0.848	0.975	0.957	0.873	0.929	0.965	0.966	15.03 M	60.7 MB

The bold values indicates the best performance.

of U-Net++ falls between our PL-Net† and PL-Net, with an IoU of 85.56% and a Dice of 91.59%. Across five cross-validation experiments, standard PL-Net showed higher stability than PL-Net†, with a 14% reduction in standard deviation. Although the Dice score of Att R2U-Net is higher than that of PL-Net, its overall performance and stability are slightly inferior. Notably, both PL-Net and BiO-Net use BSC, but our method shows better overall performance. With a smaller PL-Net† size, almost the same IoU and Dice scores as BiO-Net (t = 3, INT) can be achieved.

TN-SCUI and ACDC datasets. The boundary of the TN-SCUI dataset is blurred compared to other datasets, and we found that methods including CNN may obtain better experimental results in this case. As shown in Table 5, even lightweight approaches like PL-Net† can achieve performance similar to Swin-UNet. UNeXt-L, a hybrid network based on CNN and MLP, has the smallest model size, but its segmentation performance is inferior to baseline methods. Our analysis shows that this is because the method has fewer learnable parameters and cannot make good use of the learned features. In the ACDC dataset (Table 6), we demonstrate the segmentation performance of different methods on different classes. The target area of the myocardium (Myo) is ringed between the left atrium (LV) and right atrium (RV) and is relatively small overall. The segmentation

accuracy of different methods on this category tends to be lower than that of the other two categories. Our PL-Net achieves the highest IoU and Dice scores. Although TransUNet and EANet can achieve better average segmentation performance, their model size is increased by 6 times, making them more complex and requiring more computing resources than our proposed method. Additionally, the experimental results of PL-Net on the ACDC dataset show that our method is also suitable for multi-category segmentation tasks.

The above quantitative comparison demonstrates that our proposed network can be applied to different segmentation tasks, which can include different modalities and categories. Even for images with blurred boundaries, PL-Net can produce good segmentation results. Although the overall segmentation performance of PL-Net† is not as good as that of standard PL-Net, its smaller parameters and model size will promote its application in memory-constrained environments. Additionally, other U-Net variants, which are oriented towards functional optimization or structural optimization, can improve the segmentation performance of the original U-Net to some extent, but the increased computational cost is a difficult problem to avoid. As PL-Net is a progressive learning framework, it achieves a good trade-off between segmentation performance and parameters.

TABLE 4 Performance comparison with SOTA methods on Nuclei dataset. Red, Green, and Blue indicate the best, second-best, and third-best performance. For the original implementation methods, we report mean ± standard deviation.

Network	Nuclei dataset				#Params	Model size
	Acc	IoU	Dice	Sens		
PraNet Fan et al. (2020)	95.59	71.08	81.03	80.62	—	—
Channel-UNet Chen et al. (2019)	96.27	79.75	87.55	90.70	—	—
ResUNet Diakogiannis et al. (2020)	97.05	82.44	89.91	90.00	—	—
Double U-Net Jha et al. (2020)	—	84.07	91.33	64.07	27.94 M	112 MB
TransAttUnet D Chen et al. (2021a)	97.37	84.62	91.34	91.86	—	—
TransAttUnet R Chen et al. (2021a)	97.46	84.98	91.62	91.85	—	—
TransUNet Chen et al. (2021b)	97.84	85.21	91.69	91.62	100.4 M	401 MB
FATNet Wu et al. (2022)	98.11	85.24	91.69	91.73	27.43 M	109 MB
U-Net Ronneberger et al. (2015)	97.84 ± 0.24	85.68 ± 1.40	91.90 ± 1.00	92.61 ± 0.52	29.59 M	118 MB
U-Net++ Zhou et al. (2018)	97.87 ± 0.22	85.91 ± 1.35	92.06 ± 1.00	92.48 ± 1.08	34.48 M	138 MB
FCANet Cheng et al. (2020)	97.68 ± 0.31	84.87 ± 1.39	91.33 ± 1.09	91.70 ± 1.50	59.97 M	241 MB
Att U-Net Oktay et al. (2018)	97.84 ± 0.18	85.46 ± 1.20	91.77 ± 0.88	91.93 ± 0.66	30.42 M	121 MB
R2U-Net Alom et al. (2019)	97.93 ± 0.18	85.68 ± 1.26	91.89 ± 0.92	92.28 ± 1.20	91.61 M	366 MB
Att R2U-Net Alom et al. (2019)	97.76 ± 0.34	85.86 ± 1.04	92.15 ± 0.92	92.51 ± 1.46	92.11 M	368 MB
BiO-Net (t = 3) Xiang et al. (2020)	97.81 ± 0.22	85.09 ± 1.42	91.53 ± 1.04	91.99 ± 0.72	14.30 M	57.7 MB
BiO-Net (t = 3, INT) Xiang et al. (2020)	97.84 ± 0.20	85.31 ± 1.27	91.68 ± 0.93	91.94 ± 0.76	14.30 M	57.7 MB
UNeXt-L Valanarasu and Patel (2022)	97.43 ± 0.15	81.26 ± 1.46	88.75 ± 1.31	88.71 ± 1.65	3.99 M	15.2 MB
PL-Net† (Our)	97.79 ± 0.22	85.23 ± 1.39	91.60 ± 1.03	91.79 ± 0.59	3.77 M	15.6 MB
PL-Net (Our)	97.96 ± 0.16	86.14 ± 1.20	92.13 ± 0.88	92.12 ± 1.11	15.03 M	60.7 MB

The bold values indicates the best performance.

TABLE 5 Performance comparison with SOTA methods on TN-SCUI datasets. Red, Green, and Blue indicate the best, second-best, and third-best performance.

Network	TN-SCUI datasets		#Params (M)	Model size (MB)
	IoU	Dice		
U-Net Ronneberger et al. (2015)	0.718	0.806	29.59	118
SegNet Badrinarayanan et al. (2017)	0.726	0.819	17.94	71.8
FATNet Wu et al. (2022)	0.751	0.842	27.43	109
Swin-UNet Cao et al. (2021)	0.744	0.835	25.86	105
TransUNet Chen et al. (2021b)	0.746	0.837	88.87	401
EANet Wang et al. (2022)	0.751	0.839	47.07	118
UNeXt-L Valanarasu and Patel (2022)	0.693	0.794	3.99	15.2
PL-Net† (Our)	0.742	0.830	3.77	15.6
PL-Net (Our)	0.767	0.851	15.03	60.7

The bold values indicates the best performance.

4.4.2 Qualitative comparison

To better understand the excellent performance of our method, we present example results of PL-Net and several other methods in Figure 4 and Figure 5. As shown, our PL-Net and PL-Net† can handle different types of targets and produce accurate segmentation results.

TABLE 6 Performance comparison with SOTA methods on ACDC datasets. Red, Green, and Blue indicate the best, second-best, and third-best performance.

Network	ACDC datasets				#Params (M)	Model size (MB)
	RV	Myo	LV	Average		
U-Net Ronneberger et al. (2015)	0.743 (0.792)	0.717 (0.812)	0.861 (0.897)	0.774 (0.834)	29.59	118
SegNet Badrinarayanan et al. (2017)	0.738 (0.790)	0.720 (0.817)	0.864 (0.902)	0.774 (0.836)	17.94	71.8
FATNet Wu et al. (2022)	0.743 (0.799)	0.702 (0.805)	0.859 (0.899)	0.768 (0.834)	27.43	109
Swin-UNet Cao et al. (2021)	0.754 (0.805)	0.722 (0.820)	0.865 (0.903)	0.780 (0.843)	25.86	105
TransUNet Chen et al. (2021b)	0.750 (0.800)	0.715 (0.812)	0.872 (0.905)	0.779 (0.839)	88.87	401
EANet Wang et al. (2022)	0.742 (0.791)	0.732 (0.825)	0.864 (0.902)	0.779 (0.839)	47.07	118
UNeXt-L Valanarasu and Patel, (2022)	0.719 (0.779)	0.675 (0.810)	0.840 (0.882)	0.745 (0.824)	3.99	15.2
PL-Net† (Our)	0.723 (0.778)	0.692 (0.796)	0.845 (0.887)	0.753 (0.820)	3.77	15.6
PL-Net (Our)	0.761 (0.807)	0.738 (0.828)	0.872 (0.907)	0.790 (0.847)	15.03	60.7

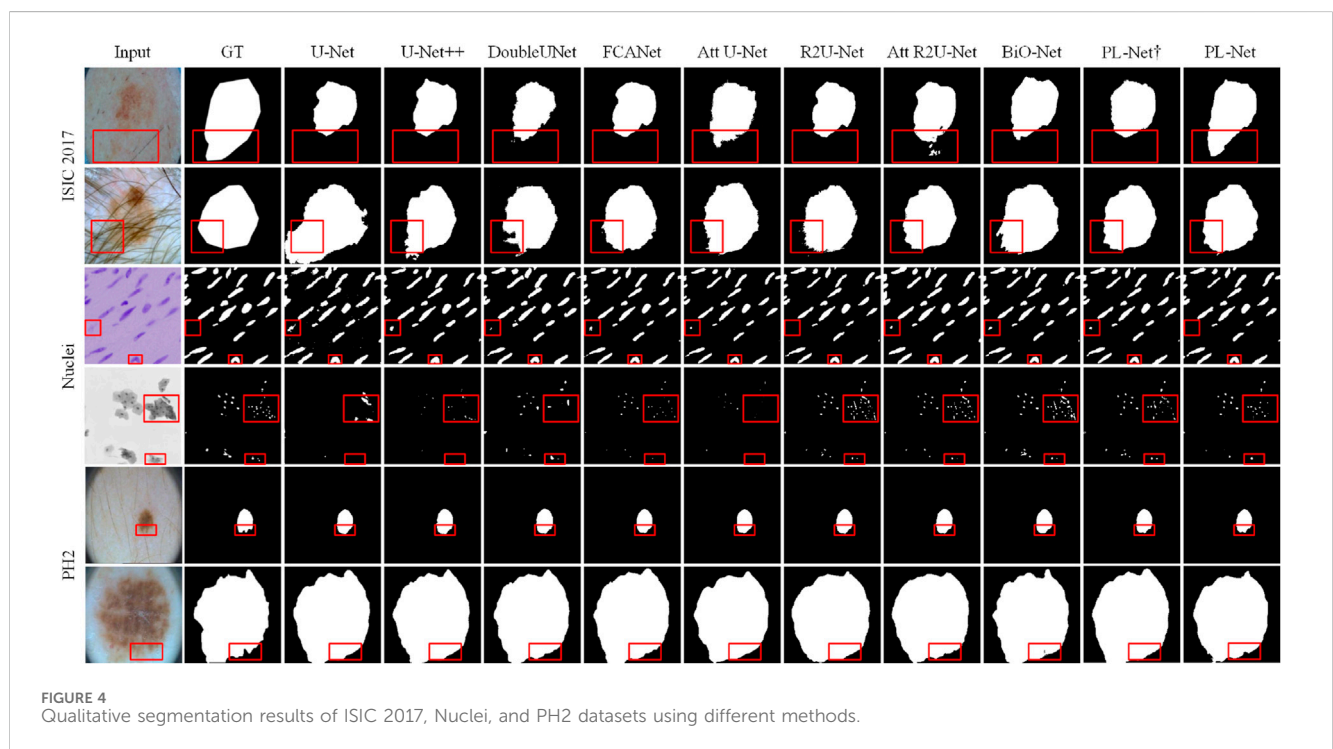The bold values indicates the best performance.

FIGURE 4
Qualitative segmentation results of ISIC 2017, Nuclei, and PH2 datasets using different methods.

The first and second rows of Figure 4 respectively show the segmentation results of an ambiguous target area and a small amount of occlusion (hair). As observed, although the results produced by PL-Net are not as accurate, our method is still effective for areas with ambiguous targets compared to other methods. When segmenting occluded images, other models either tend to divide boundaries incorrectly or mistake masked areas as target areas. The segmentation target of the image in the third row is clear, and relatively accurate segmentation results can be produced through other methods. However, for the content marked in the red box, most methods mistake interfering pixels for target pixels for segmentation, and better results are produced through our method compared to other methods. The fourth row shows the performance of different models for targets consisting of tiny targets and dispersed structures. As observed, U-Net and Att U-Net either recognize the saliency area as the target area or lose the target area, resulting in poor segmentation results. The fifth and sixth rows show the segmentation results of different methods for smaller and larger targets. As seen, our model makes a good decision on the boundary of the small target, while the area marked in the red box cannot be segmented well by other models. Compared to the fifth row, the lesion area shown in the sixth row covers almost the entire image. Although more accurate segmentation results can be produced through other methods, our PL-Net produces more perfect results as far as the area marked in the red box is concerned.

Figure 5 presents qualitative comparison results of different methods on the TN-SCUI and ACDC datasets. From the

FIGURE 5
Qualitative segmentation results of TN-SCUI and ACDC datasets using different methods.
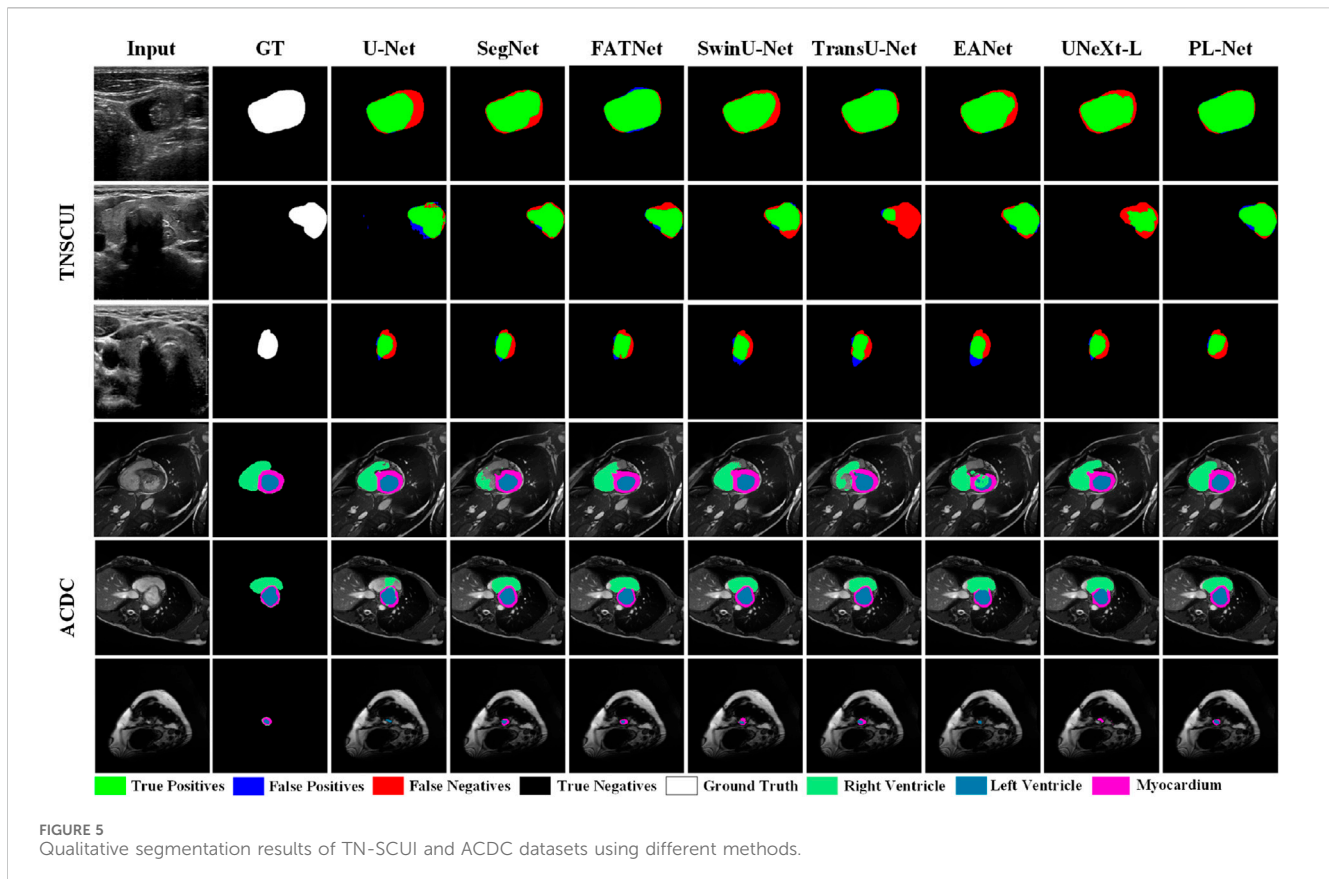
experimental results in the first two rows of the TN-SCUI dataset, PL-Net has a larger true positive area compared to other methods and is more accurate in lesion boundary segmentation. The third row shows an example where different methods perform poorly. Although there is a certain difference between our segmentation results and the ground truth, the false positive area is significantly lower than that of other methods, which is particularly important in medical image analysis. We highlighted different targets in the ACDC dataset using different colors, and the experimental results in the fourth-row show that SegNet, TransUNet, and EANet have poor segmentation results and incomplete segmentation of the RV area. In the example image in the fifth row, the Myocardium area accounts for a relatively small proportion, and FATNet, EANet, and UNeXt do not correctly segment the ring area, while PL-Net clearly segments the Myocardium area. The experimental results in the last row demonstrate the advantage of PL-Net in segmenting small targets. Although U-Net, EANet, and UNeXt segment the target area, their category definitions are inaccurate. These experimental results cover different situations in medical images, including large, medium, and small lesions, as well as targets of different categories. These results indicate that PL-Net has good generalization ability and can handle different types of medical image semantic segmentation problems.

In addition to the visualization results mentioned above, we present the features learned by different "stages" and "steps" of PL-Net in the form of a heat map, as shown in Figure 6. During internal progressive learning (i.e., "Step1″" and "Step2″"), the shallower

"Step1″" tends to focus on coarse-grained semantic information first, such as the outline of hair or lesions. As the network depth increases, "Step2″" gives less weight to texture features and focuses more on fine-grained semantics. PL-Net captures semantic information from coarse to fine granularity at different "stages" using internal progressive learning and does not introduce additional parameters compared to other approaches that replace deeper encoders. Through the visualization results of different "stages," we observe that the heat value of "Stage2″" is higher than that of "Stage1″" at the same position (i.e., the corresponding weight value is larger), which benefits from the fusion of coarse-grained and fine-grained information of the two stages. In addition, "Fusion" represents the feature map of the last convolutional layer after the two-stage fusion, with a distribution of thermal values similar to that of the ground truth and masked from irrelevant background regions.

4.4.3 Expanding to 3D medical image segmentation

In this section, we will detail how to effectively apply the proposed progressive strategy to 3D medical image segmentation tasks. To validate the effectiveness of this strategy, we chose 3D U-Net as the baseline network and conducted preliminary experimental validation on the standard prostate MRI segmentation dataset, PROMISE 2012 Litjens et al. (2014). This dataset contains 50 MRI volumes, which we split into a training set and a test set in a 3:2 ratio.

In the 3D U-Net Çiçek et al. (2016) structure, we employed basic blocks consisting of two $3 \times 3 \times 3$ convolutions (excluding batch
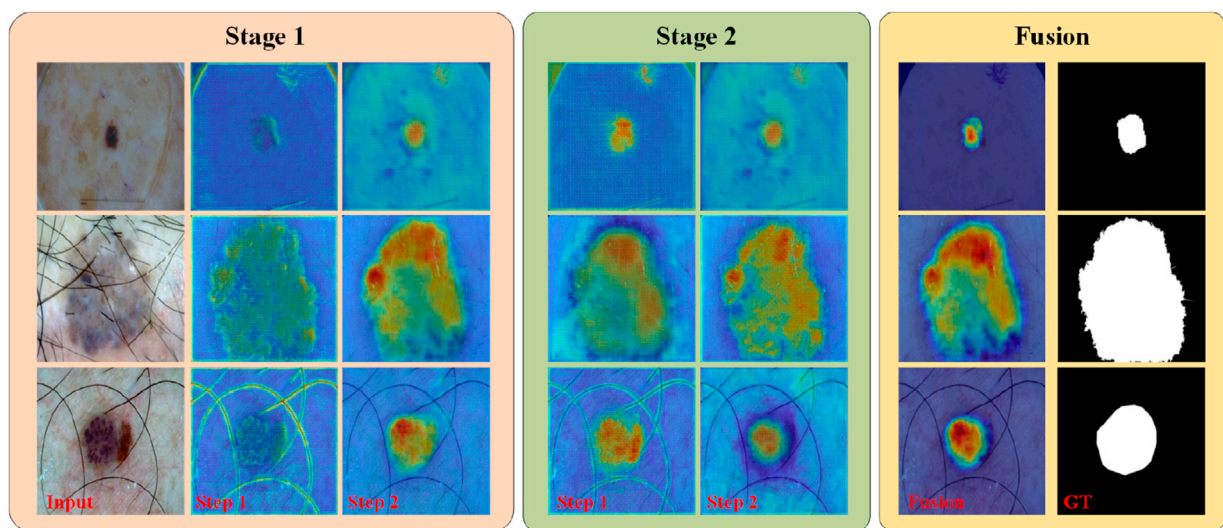
FIGURE 6
PL-Net's heat map of different "stages" and "steps" on the ISIC2017 dataset.

normalization and activation functions here). The encoder part includes four such basic blocks, each followed by a downsampling operation to progressively reduce the spatial dimensions of the feature maps. The decoder part restores the feature space through four upsampling operations, each also followed by a basic block.

To integrate the progressive learning strategy into the 3D U-Net, we converted the two $3 \times 3 \times 3$ convolutions in the basic block into an internal progressive learning process, where each convolution layer is considered a "step." In the second "step," we introduced backward skip connections to fuse features of different levels at the same scale. This design allows us to effectively incorporate the internal progressive learning strategy without altering the original 3D U-Net's basic structure.

Next, we regarded the above network as the first "stage" of external progressive learning. To construct the second "stage," we added a downsampling layer and a basic block at the end of the encoder, and an upsampling layer and a basic block at the beginning of the decoder. Similar to the design concept of PL-Net, we built the second "stage" by reusing the network structure from the first stage along with the newly added basic blocks. In the second stage, we used skip connections to effectively fuse the coarse-grained information from the first stage with the fine-grained features of the second stage. Through these steps, we extended the original 3D U-Net into a progressive learning network with two "steps" and two "stages."

As shown in Table 7, we compared the experimental results of the original 3D U-Net with those after introducing the progressive learning strategy. The data indicates that introducing only the internal progressive learning strategy improved the Dice score by 0.6% compared to the baseline 3D U-Net. When both progressive learning strategies were applied, the Dice score improvement was even more significant, reaching 1.36%. These preliminary experimental results fully demonstrate the effectiveness and practicality of our proposed method. Encouraged by these positive findings, we plan to further explore the potential

TABLE 7 Extended experiments on the application of progressive learning strategies to 3D U-Net.

Method	PROMISE 2012 dataset	
	IoU (%)	Dice (%)
3D U-Net Çiçek et al. (2016)	56.84	72.48
3D U-Net + IPL	57.58	73.08
3D U-Net + IPL + EPL	**58.53**	**73.84**

The bold values indicates the best performance.

performance of progressive learning strategies in a broader range of 3D medical image segmentation tasks in the future.

5 Discussion

U-Net has been widely used as a benchmark model for medical image segmentation due to its simple and easily modifiable structure. Most of its variant approaches enhance segmentation performance by adding functional modules (e.g., attention module) or modifying its original structure (e.g., residual, and densely connected structures) in the feed-forward process. In this paper, we adopt an alternative approach by recognizing that coarse-grained and fine-grained discriminative information naturally exists at different stages of the network, which can be learned incrementally, similar to how humans learn through shallow and deep network structures. Based on this intuition, we design a framework with internal and external progressive learning strategies, called PL-Net. Internal progressive learning strategies are used to mine semantic information at different granularities, while external progressive learning strategies further refine segmentation details based on the features learned in the previous training phase.

Researchers have proposed numerous network architectures based on U-Net to address various medical image semantic segmentation problems. However, some approaches that add functional modules (such as FCANet and Att U-Net) do not consistently improve

performance across different datasets. Our experimental results demonstrate that while FCANet improves IoU by 4% over vanilla U-Net on the ISIC2017 dataset, it degrades performance by 0.81% on the Nuclei dataset, indicating that performance variation is related to the type, size, and complexity of the dataset. Our proposed PL-Net achieves consistent performance improvements over vanilla U-Net on five datasets without adding new functional modules or structural modifications and remains competitive with state-of-the-art network frameworks (EANet and ResGANet). Moreover, PL-Net has lower computational overhead and fewer parameters, resulting in a model size reduction of 3.8 times and 6.6 times compared to widely used nnUNet and TransUNet, respectively. We also provide PL-Net† with a smaller number of parameters, which can offer options for different medical imaging scenarios, although the decrease in the number of parameters results in reduced segmentation accuracy. Our method can run on a GPU with limited memory, reducing the complex configuration and tedious preprocessing steps of nnUNet. In other words, designing such a network is crucial to translate medical imaging from the laboratory to clinical practice.

On the other hand, similar to most existing state-of-the-art methods, our proposed segmentation network still has limitations in handling cases with complex boundaries and small targets. As shown in the first row of Figure 4, when the boundary between the skin lesion and the background region is difficult to distinguish, our method and other approaches fail to accurately delineate the boundary. As shown in the third row of Figure 5, PL-Net's segmentation performance is lower when the target region is very small. However, in these cases, our method is closest to the ground truth, and the segmentation results are still better than those of other competitors. From the experimental results in Table 1, we found that the best results were obtained by performing three internal progressive learning experiments on the large-scale TN-SCUI dataset, indicating the necessity of setting different internal progressive learning strategies. Finally, we believe that introducing robust functional modules may further improve the segmentation performance of PL-Net, and we will explore this in future work. The ideas proposed in this paper mainly provide inspiration for researchers who are committed to designing feature representations to improve convolutional neural networks.

6 Conclusion

In this study, we propose a new variant of U-Net called PL-Net for 2D medical image segmentation, which mainly consists of internal and external progressive learning strategies. Compared to U-Net methods that optimize functional or structural aspects, our PL-Net achieves consistent performance improvements without additional trainable parameters. We provide both a standard PL-Net (15.03 M) and a smaller version, PL-Net† (3.77 M), to address different medical image segmentation scenarios in real-world situations. We conduct comprehensive experiments on five public medical image datasets, and the results show that PL-Net can improve the segmentation IoU of the baseline network by 0.46%–4.9%, demonstrating high competitiveness with other state-of-the-art methods.

Although our proposed method has shown promising results, it still has some limitations that need to be further addressed in future

research: 1) Impact of data size: Exploring the parameter settings of internal and external progressive learning under different data sizes will help researchers understand the potential of the model under different scales of data. In the future, we will further explore the performance of PL-Net on larger datasets. 2) Due to the limitations of computing power and data, our method mainly focuses on 2D medical image segmentation. This article has initially demonstrated the feasibility of the progressive learning strategy in 3D medical image segmentation. In the future, we will extend PL-Net to more advanced 3D medical image segmentation frameworks to further enhance its capabilities in 3D medical image segmentation. 3) Design of functional modules: How to design functional modules suitable for PL-Net to improve segmentation performance while maintaining a concise framework is also a topic for further research in the future.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

KM: Conceptualization, Methodology, Project administration, Writing–review and editing. RL: Writing–review and editing, Conceptualization, Investigation, Methodology. JC: Writing–original draft, Conceptualization, Software. DH: Writing–review and editing, Formal Analysis. ZS: Writing–review and editing, Validation. ZL: Writing–review and editing, Visualization.

Funding

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al-Masni, M. A., Al-Antari, M. A., Choi, M.-T., Han, S.-M., and Kim, T.-S. (2018). Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Comput. methods programs Biomed.* 162, 221–231. doi:10.1016/j.cmpb.2018.05.027

Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., and Asari, V. K. (2019). Recurrent residual u-net for medical image segmentation. *J. Med. imaging* 6, 1–014006. doi:10.1117/1.jmi.6.1.014006

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. pattern analysis Mach. Intell.* 39, 2481–2495. doi:10.1109/tpami.2016.2644615

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., et al. (2018). Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans. Med. imaging* 37, 2514–2525. doi:10.1109/tmi.2018.2837502

Berseth, M. (2017) *Isic 2017 - skin lesion analysis towards melanoma detection. ArXiv* abs/1703.00523.

Caicedo, J. C., Goodman, A., Karhohs, K. W., Cimini, B. A., Ackerman, J., Haghighi, M., et al. (2019). Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nat. methods* 16, 1247–1253. doi:10.1038/s41592-019-0612-7

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). "Swin-unet: unet-like pure transformer for medical image segmentation," in *ECCV workshops*.

Chen, B., Liu, Y., Zhang, Z., Lu, G., and Kong, A. W.-K. (2021a). Transattunet: multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* 8, 55–68. doi:10.1109/tetci.2023.3309626

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021b) *Transunet: transformers make strong encoders for medical image segmentation.* arXiv preprint arXiv:2102.04306.

Chen, Y., Wang, K., Liao, X., Qian, Y., Wang, Q., Yuan, Z., et al. (2019). Channel-unet: a spatial channel-wise convolutional neural network for liver and tumors segmentation. *Front. Genet.* 10, 1110. doi:10.3389/fgene.2019.01110

Cheng, J., Tian, S., Yu, L., Gao, C., Kang, X., Ma, X., et al. (2022). Resganet: residual group attention network for medical image classification and segmentation. *Med. Image Anal.* 76, 102313. doi:10.1016/j.media.2021.102313

Cheng, J., Tian, S., Yu, L., Lu, H., and Lv, X. (2020). Fully convolutional attention network for biomedical image segmentation. *Artif. Intell. Med.* 107, 101899. doi:10.1016/j.artmed.2020.101899

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3d u-net: learning dense volumetric segmentation from sparse annotation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, 424–432. doi:10.1007/978-3-319-46723-8_49

Diakogiannis, F. I., Waldner, F., Caccetta, P., and Wu, C. (2020). Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogrammetry Remote Sens.* 162, 94–114. doi:10.1016/j.isprsjprs.2020.01.013

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020) *An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.*

Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., et al. (2020). Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. *IEEE Trans. Med. imaging* 39, 3679–3690. doi:10.1109/tmi.2020.3002417

Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., et al. (2020). "Pranet: parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention* (Berlin, Germany: Springer), 263–273.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., et al. (2019). Ce-net: context encoder network for 2d medical image segmentation. *IEEE Trans. Med. imaging* 38, 2281–2292. doi:10.1109/tmi.2019.2903562

Hasan, M. K., Dahal, L., Samarakoon, P. N., Tushar, F. I., and Martí, R. (2020). Dsnet: automatic dermoscopic skin lesion segmentation. *Comput. Biol. Med.* 120, 103738. doi:10.1016/j.compbiomed.2020.103738

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, Hawaii, USA, 21-26 July 2017, 7132–7141. doi:10.1109/cvpr.2018.00745

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. methods* 18, 203–211. doi:10.1038/s41592-020-01008-z

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y. (2017). "The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, Long Beach, CA, USA, June 16 2019 to June 17 2019, 11–19. doi:10.1109/cvprw.2017.156

Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., and Johansen, H. D. (2020). "Doubleu-net: a deep convolutional neural network for medical image segmentation," in 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS) (IEEE), Rochester, MN, USA, July 28 2020 to July 30 2020, 558–564. doi:10.1109/cbms49503.2020.00111

Kaul, C., Manandhar, S., and Pears, N. (2019). "Focusnet: an attention-based fully convolutional network for medical image segmentation," in 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019) (IEEE), Venice, Italy, 8-11 April 2019, 455–458. doi:10.1109/isbi.2019.8759477

Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., et al. (2020). Skin lesion segmentation via generative adversarial networks with dual discriminators. *Med. Image Anal.* 64, 101716. doi:10.1016/j.media.2020.101716

Litjens, G., Toth, R., Van De Ven, W., Hoeks, C., Kerkstra, S., Van Ginneken, B., et al. (2014). Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Med. image Anal.* 18, 359–373. doi:10.1016/j.media.2013.12.002

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, June 7 2015 to June 12 2015, 3431–3440. doi:10.1109/cvpr.2015.7298965

Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., and Rozeira, J. (2013). "PH² - a dermoscopic image database for research and benchmarking," in 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (IEEE), Osaka, Japan, 3-7 July 2013, 5437–5440. doi:10.1109/embc.2013.6610779

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M. J., Heinrich, M. P., Misawa, K., et al. (2018) *Attention u-net: learning where to look for the pancreas. ArXiv abs/1804.03999.*

Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., and Romero, E. (2015). An open access thyroid ultrasound image database. *10th Int. symposium Med. Inf. Process. analysis (SPIE)* 9287, 188–193. doi:10.1117/12.2073532

Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., and Soler, L. (2021). "U-net transformer: self and cross attention for medical image segmentation," in Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, 267–276. doi:10.1007/978-3-030-87589-3_28

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. *Fronrs Neurosci. abs/1505*, 04597. doi:10.1007/978-3-319-24574-4_28

Roy, A. G., Navab, N., and Wachinger, C. (2018). "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, 421–429. doi:10.1007/978-3-030-00928-1_48

Sun, J., Darbehani, F., Zaidi, M., and Wang, B. (2020). "Saunet: shape attentive u-net for interpretable medical image segmentation," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, 797–806. doi:10.1007/978-3-030-59719-1_77

Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., et al. (2015). Discriminative dictionary learning for abdominal multi-organ segmentation. *Med. image Anal.* 23, 92–104. doi:10.1016/j.media.2015.04.015

Valanarasu, J. M. J., and Patel, V. M. (2022). "Unext: mlp-based rapid medical image segmentation network," in International conference on medical image computing and computer-assisted intervention, Vancouver, BC, Canada, 8-12 October 2022, 23–33. doi:10.1007/978-3-031-16443-9_3

Wang, K., Zhang, X., Zhang, X., Lu, Y., Huang, S., and Yang, D. (2022). Eanet: iterative edge attention network for medical image segmentation. *Pattern Recognit.* 127, 108636. doi:10.1016/j.patcog.2022.108636

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., and Wen, Z. (2022). Fat-net: feature adaptive transformers for automated skin lesion segmentation. *Med. image Anal.* 76, 102327. doi:10.1016/j.media.2021.102327

Xiang, T., Zhang, C., Liu, D., Song, Y., Huang, H., and Cai, W. (2020). "Bio-net: learning recurrent bi-directional connections for encoder-decoder architecture," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, 74–84. doi:10.1007/978-3-030-59710-8_8

Xu, J., Chutatape, O., Sung, E., Zheng, C., and Kuan, P. C. T. (2007). Optic disk feature extraction via modified deformable model technique for glaucoma analysis. *Pattern Recognit.* 40, 2063–2076. doi:10.1016/j.patcog.2006.10.015

Zhou, Y., Onder, O. F., Dou, Q., Tsougenis, E., Chen, H., and Heng, P.-A. (2019). "Cia-net: robust nuclei instance segmentation with contour-aware information aggregation," in Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, 682–693. doi:10.1007/978-3-030-20351-1_53

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). "Unet++: a nested u-net architecture for medical image segmentation," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, 3–11. doi:10.1007/978-3-030-00889-5_1