Computers and Electrical Engineering 99 (2022) 107777

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

LAEDNet: A Lightweight Attention Encoder–Decoder Network for ultrasound medical image segmentation ☆

Quan Zhou [a,*], Qianwen Wang [a], Yunchao Bao [a], Lingjun Kong [b], Xin Jin [c], Weihua Ou [d]

[a] National Engineering Research Center of Communications and Networking, Nanjing University of Posts and Telecommunications, Nanjing, China
[b] Faculty of Network and Telecommunication Engineering, Jinling Institute of Technology, Nanjing, China
[c] Department of Computer Science and Technology, Beijing Electronic Science and Technology Institute, Beijing, China
[d] School of Big Data and Computer Science, Guizhou Normal University, Guiyang, China

ARTICLE INFO

Keywords:
Lightweight network
Medical ultrasound image segmentation
Encoder–decoder network
Visual attention
EfficientNet

ABSTRACT

Automatic ultrasound image segmentation plays an important role in early diagnosis of human diseases. This paper introduces a novel and efficient encoder–decoder network, called Lightweight Attention Encoder–Decoder Network (LAEDNet), for automatic ultrasound image segmentation. In contrast to previous encoder–decoder networks that involve complicated architecture with numerous parameters, our LAEDNet adopts lightweight version of EfficientNet as encoder. On the other hand, a Lightweight Residual Squeeze-and-Excitation (LRSE) block is employed in decoder. To achieve trade-off between segmentation accuracy and implementing efficiency, we also present a family of models, from light to heavy (denoted as LAEDNet-S, LAEDNet-M, and LAEDNet-L, respectively), with varying lightweight version of EfficientNet backbones. To evaluate LAEDNet, we have conducted extensive experiments on Brachial Plexus Dataset (BP), Breast Ultrasound Images Dataset (BUSI), and Head Circumference Ultrasound Images Dataset (HCUS), where ultrasound images are suffered from high noise, blurred borders and low contrast. The experiments show that, compared with U-Net and its variants, e.g., M-Net, U-Net++ and TransUNet, our LAEDNet achieves better results in terms of Dice Coefficient (DSC) and running speed. Particularly, LAEDNet-M only has 10.75M model parameters with 40.7 FPS, yet obtaining 73.0%, 73.8% and 91.3% DSC on BP, BUSI and HCUS datasets, respectively.

1. Introduction

Compared with Magnetic Resonance Imaging (MRI) and Computer Tomography (CT), Ultrasound (US) for diagnosing and analyzing the internal structure of the human body is widely used in the medical field because of its portability, low cost and real-time nature [1]. In particular, US is the most common application technique to diagnose brachial plexus neurons and lesions of breast tumors [2]. Image segmentation is widely used to obtain prominent parts and Regions of Interest (ROI) in medical images, which is conducive to disease diagnosis and pathological follow-up processing. However, due to the inherent characteristics of noise, shadow, speckle, low contrast and blurred edges of ultrasound images, it is a challenge to segment targets from ultrasound images.

Segmentation for ultrasound images has been widely studied in the community of computer vision. Potočnik et al. [3] proposed a hybrid method of spatial constrained kernel fuzzy clustering and edge-based active contour for ultrasonic image segmentation,

☆ This paper is for regular issues of CAEE. Reviews were processed and recommended for publication by Co-Editor in Chief Prof Huimin Lu.

* Corresponding author.
E-mail address: quan.zhou@njupt.edu.cn (Q. Zhou).

https://doi.org/10.1016/j.compeleceng.2022.107777
Received 9 July 2021; Received in revised form 27 January 2022; Accepted 29 January 2022
Available online 2 March 2022

where a distance regularized level set function is used. Giraldo et al. [4] introduced a method based on Bayesian shape model, and the shape prior is initialized by calculating the average shape of the neural structure in the training set. These early attempts rely on prior information to extract features, yet have limited representation ability, leading to the incorrect recognition of lesion regions in complex background.

Recently, the convolutional neural networks (CNNs) have achieved remarkable progress for the task of ultrasound medical image segmentation [5–10]. Among these networks, the symmetric encoder–decoder architecture like U-Net [11] is the mainstream model architecture for medical image segmentation. Recently, a series of U-Net variants [7,12–14] have been proposed to improve performance. For example, M-Net [13] is an effective modification of U-Net [11], which uses multi-scale inputs and outputs from different levels to monitor the extracted feature maps intensively. Abraham et al. [7] improved U-Net [11] and M-Net [13] using average pooling instead of maximum pooling to segment brachial plexus. U-Net++[14] is another powerful encoder–decoder architecture based on U-Net [11], adopting dense connections between each stage of encoder and decoder. TransUNet [15] adopts an architecture which merits both Transformers [16] and U-Net [11] to enhance finer details by recovering localized spatial information. Although these advances have achieved higher segmentation performance, due to the huge amount of model parameters, they are at the cost of expensive computing and low implementing efficiency. However, the execution efficiency of the medical assistance system is crucial in clinical practice. A lightweight network that with a small model size is able to achieve real-time segmentation in a timely fashion (e.g., at least 24 FPS inference speed,[1]) which is essential for doctors to quickly locate the target position in clinical diagnosis, thus ensuring the smooth progress of treatment, such as anesthetic injection and ultrasonic puncture operation.

In this paper, our goal is to solve the trade-off of accuracy and efficiency as a whole for ultrasound medical image segmentation, rather than just stand on one side. More specifically, our LAEDNet follows the encoder–decoder architecture that is commonly-used for image segmentation tasks. However, medical image segmentation is more challenging since ultrasound images often contains unclear target boundaries and noise background. Intuitively, in the scenario of real-time medical assistance, segmenting ultrasound images not only asks for more powerful encoder backbone to abstract high-level features, but also requires lightweight decoder to integrate extracted features from encoder and recover feature resolutions. Following this principle, on one hand, the encoder of our LAEDNet adopts various lightweight version of EfficientNet [18] to enhance feature extraction, while keeping low computational costs. Unlike traditional networks (e.g.,U-Net [11], U-Net++ [14], and TransUNet [15]) that employs simple concatenation operation to fuse encoder cues, on the other hand, our decoder adopts a lightweight information integration module with the guidance of attention scheme, called LRSE, to recover feature resolutions step-by-step, where deconvolution features are coupled with corresponding encoder counterparts, leading to produce more accurate segmentation results and more smoother object contours. We evaluate our LAEDNet on three challenging datasets, Brachial Plexus Dataset (BP) [19], Breast Ultrasound Images Dataset (BUSI) [20], and Head Circumference Ultrasound Images Dataset (HCUS) [21]. The experimental results show that our method is able to obtain available trade-off in terms of segmentation accuracy and implementing efficiency. In summary, the main contributions of our paper are three-fold:

- The asymmetric architecture of LAEDNet leads to the great reduction of network parameters, which accelerates the inference process.
- We design a lightweight decoder block, LRSE, with the guidance of attention scheme, which can be well coupled with LAEDNet backbone.
- We test LAEDNet on BP [19], BUSI [20], and HCUS [21]. The comprehensive experiments demonstrate that our LAEDNet-M achieves best trade-off between segmentation accuracy and running efficiency. Particularly, LAEDNet-M only has 10.75M model parameters with 40.7 FPS, yet obtaining 73.0%, 73.8% and 91.3% DSC on BP [19], BUSI [20] and HCUS [21], respectively.

The remainder of this paper is organized as follows. We first review related work in Section 2, and then elaborate on the details of our LAEDNet in Section 3. Experimental results are demonstrated in Section 4. Results of the ablation experiments are discussed in Section 5. Finally, the conclusion and future work are given in Section 6.

2. Related work

In this section, we review the related advances for US medical image segmentation using encoder–decoder architecture. As ultrasound medical image segmentation and visual attention are most related fields to our work, we review the related approaches in these two directions.

2.1. Ultrasound medical image segmentation

The recent mainstream approaches prefer to design high accurate networks for ultrasound medical image segmentation. As a pioneer work, U-Net [11] designs a mirrored architecture for medical image segmentation. Amiri et al. [22] fine-tune the shallow layers rather than deep layers in ultrasound image segmentation. The NAS-Unet [23] is constructed based on a U-like backbone [11] to medical image segmentation through neural architecture search. This method needs to manually propose rules of searching space, thus increasing the complexity of network design. To segment regions of breast masses with different size and shape, Michal

[1] Due to the effect of persistence of vision, human visual system is not sensitive to the change of still images, when they are played at least with 24 FPS [17].

et al. [24] develop a selective kernel U-Net, integrating the merit of dilated convolution and attention module to adaptively enlarge receptive fields. To capture context information, FC-DenseNet [25] replaces each upsampling and downsampling layer of U-Net [11] for kidney segmentation in US Images. Wu et al. [26] cascade fully convolutional network for prenatal US image segmentation. In order to learn long-range spatial context of transvaginal US image segmentation, CR-Unet [27] incorporates the spatial recurrent neural network (RNN) [28] into a plain U-Net [11]. Yang et al. [29] adopts the RNN in the decoder of U-Net [11] to segment prenatal volumetric US image. In DeepNerve [30], RNN is embedded into the encoder for median nerve US image segmentation. Although these methods have made remarkable progress for ultrasound medical image segmentation, they are inevitable to slow down the inference speed because of their huge network model size, which is not suitable for assisting urgent clinical tasks in timely fashion. Conversely, our LAEDNet adopts lightweight version of EfficientNet [18] as encoder, and designs a set of lightweight attention blocks in decoder, achieving trade-off between segmentation accuracy and implementing efficiency.

Recently, there are two directions of designing a real-time segmentation network: (1) optimizing existing high-accuracy networks, such as pruning [31,32], quantization [33,34], and distillation [35–37], and (2) designing compact lightweight networks, such as LEDNet [38] and AGLNet [39]. Although there are a vast number of compact networks designed for semantic segmentation, to our best knowledge, there are few researches that study the medical imaging problems. Some early attempts [36,37,40] have started paying attention to real-time medical image segmentation. Yet there is still a dilemma that the performance tends to be damaged when the models are simplified only for faster speed. In contrast, our LAEDNet belongs to the second category, which seeks best trade-off between accuracy and efficiency for ultrasound medical image segmentation with the guidance of attention scheme.

2.2. Visual attention

The recent visual attention used in ultrasound medical image segmentation can be roughly divided into three categories: squeeze attention [41,42], self-attention [43], and transformers [16].

The first category often utilizes global average pooling to produce channel or spatial attention used to reweight original feature maps [41,42]. For example, AttentionNet [44] embeds the spatial attention refinement (SAR) block into backbone to extract more informative features for breast US images segmentation. ReAgU-Net [45] employs the attention gate (AG) to reweight feature maps obtained from shallow layers and deep layers for thyroid US image segmentation. Wang et al. [46] adopt the layer-wise attention mechanism to selectively leverage the complementary features across all scales. In contrast, the second category [43,47] encodes global context by computing correlation matrix of image elements, showing more powerful representation ability with respect to squeeze attention scheme. For instance, Xue et al. [47] not only consider channel-wise attention, but also embed long-range spatial dependencies. LEDNet [38] and AGLNet [39] adopt additional spatial attention and feature pyramid attention in decoder, yet both channel attention and spatial attention are not designed in a lightweight way. Transformer [15,16,48], derived from nature language processing, begins to show their potential for ultrasound image segmentation. For instance, TransUNet [15] attempts to combine the metric of Transformers [16] and U-Net [11] for medical image segmentation. In sprite of having powerful ability to capture global context, these attention models once again are at sacrifice of very large network size, resulting in low implementing efficiency. Unlike these approaches, our goal is to design a lightweight attention module, called LRSE, to achieve accurate segmentation outputs, while maintaining fast segmenting speed.

3. Method

3.1. Overall network architecture

The overall network architecture is shown in Fig. 1. LAEDNet also adopts the encoder–decoder architecture similar to U-Net [11], but the specific details of the encoder and the decoder are asymmetrical.

The contraction path of LAEDNet employs an efficient backbone, EfficientNet [18], as encoder to reduce model size while extract robust features. There are a total of 8 stages (denoted as gray blocks) in the encoder. The output resolution of stem stage is 1/2 of the input image. Except first stem block, let X_E^i be the output of the ith block in encoder, then the output resolution of each individual block is 1/2, 1/4, 1/8, 1/16, 1/16, 1/32 and 1/32 with respect to the size of original image. On the other hand, there are six green blocks (LRSE) in the decoder, where feature resolutions are sequentially recovered.

Let X_D^{j-1} be the output of the jth LRSE. As can be seen, the jth LRSE has two inputs, X_D^j from deeper LRSE, and X_E^i from the counterpart in encoder. Note the feature resolutions are required to be upsampled (indicated as red arrows) for exact integration in each LRSE. Finally, an 1×1 convolution is adopted to project feature maps to semantic space. Immediately below, we elaborate on the details of encoder and decoder, respectively.

3.2. Encoder

To achieve performance balance in terms of segmentation accuracy and implementing efficiency, LAEDNet employs three versions of EfficientNet [18] architecture, EfficientNet-B0, EfficientNet-B3, and EfficientNet-B7 (denoted as LAEDNet-S, LAEDNet-M, and LAEDNet-L, respectively), where the model size is scaled by the number of convolution layers. For convenient understanding, Fig. 2 takes LAEDNet-S as an example to show the detail structure of backbone. The entire backbone is composed of a stem and a series of convolution blocks, where each block includes a set of repeatably MBConv modules with different size of depthwise convolutions.

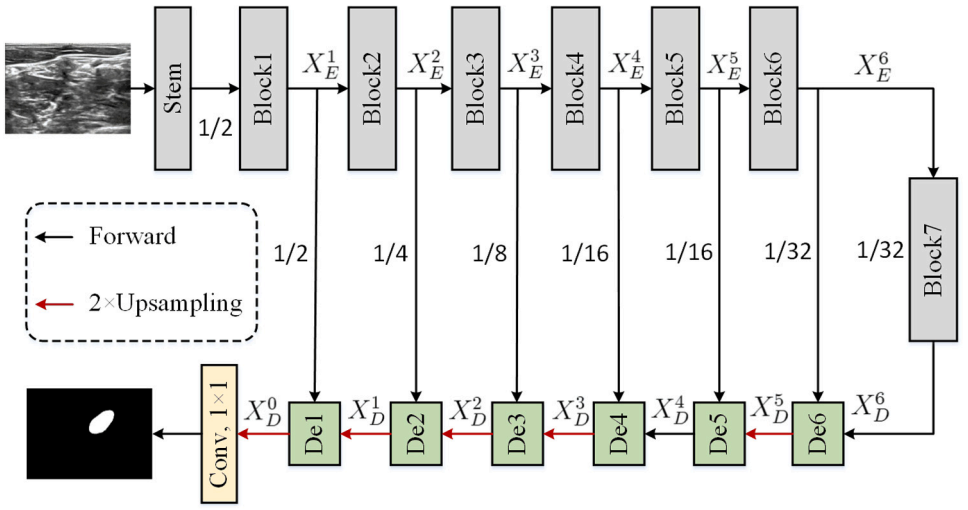

Fig. 1. The overall network architecture of LAEDNet. The gray blocks are a series of encoding modules, and the green blocks represent LRSEs. The black arrows denote information flow, and red arrows indicate 2 times upsampling operation. Note that the entire network has an asymmetrical architecture, as the detail structure in LRSEs is different with convolution stages in encoder.

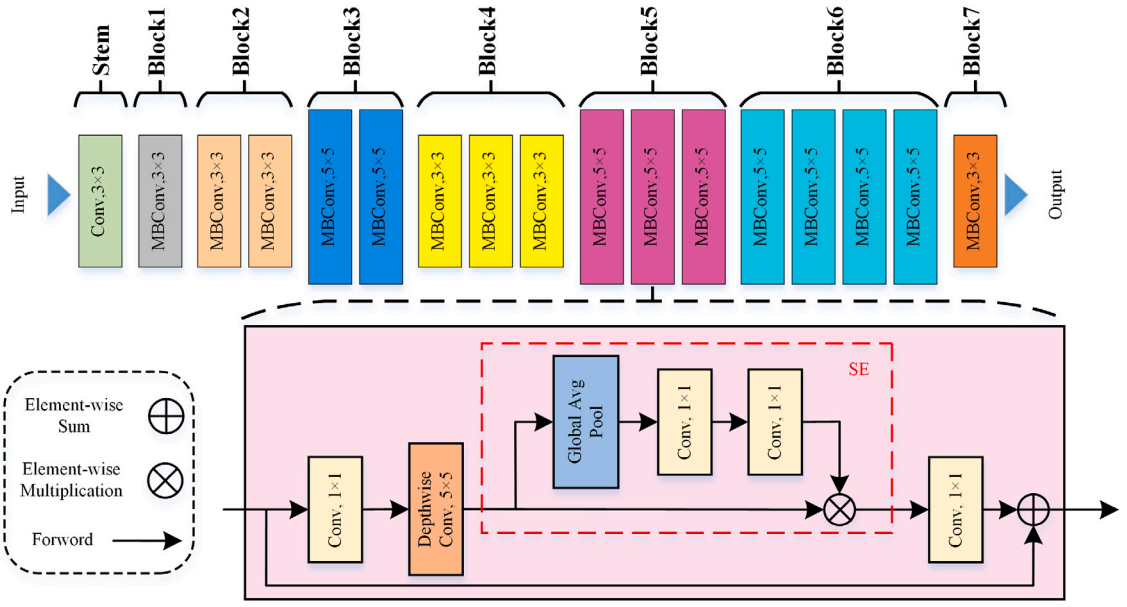

Fig. 2. The structure of encoder in LAEDNet. Different blocks are denoted by different colors. The details of the structure of MBConv are described in the purple block.

The MBConv adopts inverted residual architecture that feature channels are first expended and reweighed by squeeze-and-excitation (denoted as red dash line bounding box). Finally, the channel number of reweighed features are reduced for residual connection.

More specifically, the detailed encoder structure of LAEDNet-S, LAEDNet-M, and LAEDNet-L are exhibits in Table 1. The input size of the image is $320 \times 320 \times 3$, and the output size of each block sequentially decreases along with going deeper of entire network. All three types of encoders are stacked by a stem layer and 7 convolution blocks. Specifically, the stem block is a convolution layer using 3×3 filtering with stride 2, which reduce the dimension of the input image. The next 7 blocks consist of a set of mobile inverted bottlenecks (MBConvs) [49] with the squeeze-and-excitation (SE) [42] optimization. In different lightweight version of LAEDNet, MBConv is repeated with different times, leading to different number of convolution layers and model size. Considering LAEDNet-S as an example, MBConvs are repeated m times in each convolution block, where different block stages has different structures of MBConv. In each MBConv, an 1×1 convolution is first applied to increase the number of feature channels to (c_1,c_2), where c_1 is only used for first MBConv in each block, and the rest MBConvs adopt c_2. Thereafter, a depthwise convolution is performed with 3×3 filter kernel size. The filtering feature maps are reweighted by the forthcoming SE attention module. Finally, another 1×1

Table 1
Detailed encoder architecture of LAEDNet-S, LAEDNet-M, and LAEDNet-L.

Layer name	Output size	LAEDNet-S	LAEDNet-M	LAEDNet-L
Stem	160 × 160	[3 × 3, 32] × 1	[3 × 3, 40] × 1	[3 × 3, 64] × 1
Block1	160 × 160	$\begin{bmatrix} DW\,3\times3, 32 \\ SE \\ 1\times1, 16 \end{bmatrix} \times 1$	$\begin{bmatrix} DW\,3\times3, 40 \\ SE \\ 1\times1, 24 \end{bmatrix} \times 1$	$\begin{bmatrix} DW\,3\times3, (64, 32) \\ SE \\ 1\times1, 32 \end{bmatrix} \times 4$
Block2	80 × 80	$\begin{bmatrix} 1\times1, (96, 144) \\ DW\,3\times3, (96, 144) \\ SE \\ 1\times1, 24 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, (144, 192) \\ DW\,3\times3, (144, 192) \\ SE \\ 1\times1, 32 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, (192, 288) \\ DW\,3\times3, (192, 288) \\ SE \\ 1\times1, 48 \end{bmatrix} \times 7$
Block3	40 × 40	$\begin{bmatrix} 1\times1, (144, 240) \\ DW\,5\times5, (144, 240) \\ SE \\ 1\times1, 40 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, (192, 288) \\ DW\,5\times5, (192, 288) \\ SE \\ 1\times1, 48 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, (288, 480) \\ DW\,5\times5, (288, 480) \\ SE \\ 1\times1, 80 \end{bmatrix} \times 7$
Block4	20 × 20	$\begin{bmatrix} 1\times1, (240, 480) \\ DW\,3\times3, (240, 480) \\ SE \\ 1\times1, 80 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, (288, 576) \\ DW\,3\times3, (288, 576) \\ SE \\ 1\times1, 96 \end{bmatrix} \times 5$	$\begin{bmatrix} 1\times1, (480, 960) \\ DW\,3\times3, (480, 960) \\ SE \\ 1\times1, 160 \end{bmatrix} \times 10$
Block5	20 × 20	$\begin{bmatrix} 1\times1, (480, 672) \\ DW\,5\times5, (480, 672) \\ SE \\ 1\times1, 112 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, (576, 816) \\ DW\,5\times5, (576, 816) \\ SE \\ 1\times1, 136 \end{bmatrix} \times 5$	$\begin{bmatrix} 1\times1, (960, 1344) \\ DW\,5\times5, (960, 1344) \\ SE \\ 1\times1, 224 \end{bmatrix} \times 10$
Block6	10 × 10	$\begin{bmatrix} 1\times1, (672, 1152) \\ DW\,5\times5, (672, 1152) \\ SE \\ 1\times1, 192 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, (816, 1392) \\ DW\,5\times5, (816, 1392) \\ SE \\ 1\times1, 232 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, (1344, 2304) \\ DW\,5\times5, (1344, 2304) \\ SE \\ 1\times1, 384 \end{bmatrix} \times 13$
Block7	10 × 10	$\begin{bmatrix} 1\times1, (1152, 1920) \\ DW\,3\times3, (1152, 1920) \\ SE \\ 1\times1, 320 \end{bmatrix} \times 1$	$\begin{bmatrix} 1\times1, (1392, 2304) \\ DW\,3\times3, (1392, 2304) \\ SE \\ 1\times1, 384 \end{bmatrix} \times 1$	$\begin{bmatrix} 1\times1, (2304, 3840) \\ DW\,3\times3, (2304, 3840) \\ SE \\ 1\times1, 640 \end{bmatrix} \times 4$

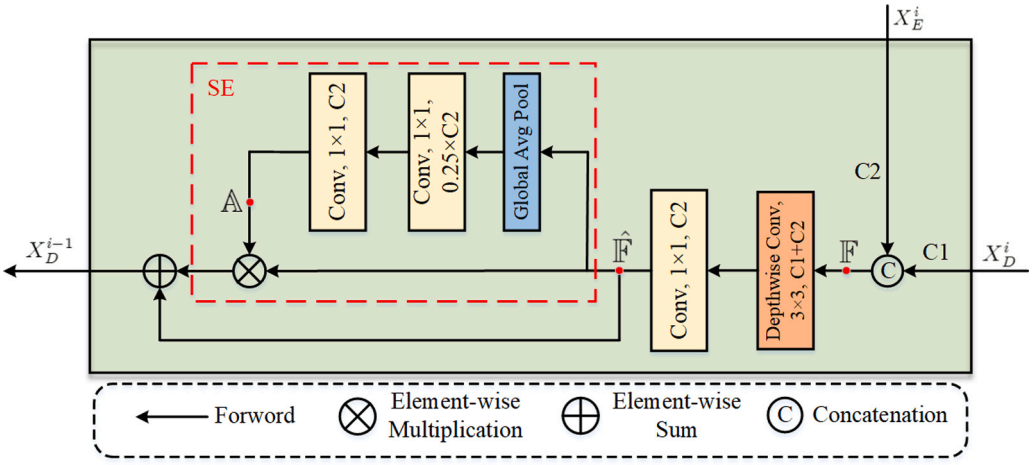

Fig. 3. Illustration of LRSE. The orange, yellow, and blue module represent the depthwise convolution, 1 × 1 convolution and global average pooling, respectively.

convolution is employed to reduce feature dimensions, where the output feature maps are fed into next MBConv. Compared with LAEDNet-S, LAEDNet-M and LAEDNet-L adopt similar architecture, yet have different number of filter feature channels and repeated times m of MBConv.

3.3. Decoder

The detailed architecture of LRSE used in decoder is depicted in Fig. 3, where LRSE makes the decoder well coupled with the encoder. For ith LRSE in decoder, it integrates the upsampled features $X_D^i \in \mathbb{R}^{W\times H\times C1}$ from previous LRSE, and the counterpart feature maps $X_E^i \in \mathbb{R}^{W\times H\times C2}$ in encoder to produce output features $X_D^{i-1} \in \mathbb{R}^{W\times H\times C2}$. Note, as shown in Fig. 1, there is no need to

upsample X_D^4 and X_D^6, as they already have the same resolution with respect to the corresponding X_E^4 and X_E^6. Specifically, we first fuse X_D^i and X_E^i by concatenation:

$$\mathbb{F} = [X_D^i, X_E^i] \tag{1}$$

where $[\cdot]$ denotes concatenated operation. Then, the produced features $\mathbb{F} \in \mathbb{R}^{W \times H \times (C1+C2)}$ are fed into a 3×3 depthwise convolution, where each feature channel is convoluted independently. To construct correlations among feature channels, an 1×1 convolution is applied, and the feature dimension is reduced at the same time. Let F_l be these lightweight convolution operations, and θ_l is the associated parameters. Then the output features $\hat{\mathbb{F}} \in \mathbb{R}^{W \times H \times C2}$ can be calculated as:

$$\hat{\mathbb{F}} = F_l(\mathbb{F}, \theta_l) \tag{2}$$

The final step of LRSE is reweighting $\hat{\mathbb{F}}$ with the guidance of SE [42] attention. To this end, an average pooling is first used to squeeze input feature $\hat{\mathbb{F}}$ into a one-dimensional vector, then two 1×1 convolutions are employed to produce a channel-wised attention map $\mathbb{A} \in \mathbb{R}^{1 \times 1 \times C2}$:

$$\mathbb{A} = F_{SE}(\mathbf{avg}(\hat{\mathbb{F}}), \theta_{SE}) \tag{3}$$

where $avg(\cdot)$ is an average pooling operation, F_{SE} indicates projection using 1×1 convolutions and θ_{SE} is the associated parameters. Finally, the output $X_D^{i-1} \in \mathbb{R}^{W \times H \times C2}$ is calculated by combined features $\hat{\mathbb{F}}$ and its reweighted counterpart:

$$X_D^{i-1} = \hat{\mathbb{F}} + \mathbb{A} \bigotimes \hat{\mathbb{F}} \tag{4}$$

where $\bigotimes$ stands for element-wise multiplication, and an identity mapping is used to leverage model training and feature reweighting.

Note SE is both used in encoder and decoder of LAEDNet, yet it has different effects. In encoder, SE is part of MBConv, and used to reweight feature maps with expanded channels. In decoder, on the contrary, SE is used to reweight integrated feature maps, derived from previous deconvolution features and the encoder counterpart.

4. Experiments

To demonstrate the effectiveness and scalability of LAEDNet, we have conducted extensive experiments for different medical segmentation scenarios: BP [19] for neck, BUSI [20] for chest, and HCUS [21] for fetal head in the womb. The experimental results show that our LAEDNet achieves best balance between segmentation accuracy and implementing efficiency on these datasets.

4.1. Dataset

BP [19] dataset (https://www.kaggle.com/c/ultrasound-nerve-segmentation) was released by Kaggle competition in 2016. It consists of 5635 ultrasound images of brachial plexus neurons from 47 patients, including 2323 images with neuron targets and 3312 images without neuron targets. For each patient, about 120 images were collected. The resolution of all original images are 420×580. In data pre-processing, we divide the entire dataset into training set and validation set according to the ratio of 9:1. Specifically, to ensure the same sample distribution between the training set and the validation set, we construct validation set using 1/10 of the images with and without neuron targets. The rest of the images are used to form training set.

BUSI [20] dataset (https://scholar.cu.edu.eg/?q=afahmy/pages/dataset) was collected in 2018 and includes breast ultrasound images of 600 women aged 25 to 75. The dataset consists of 780 images in PNG format, with an average image size of 500×500 pixels. The images in BUSI [20] are divided into three categories: normal, benign and malignant, with 133, 437 and 210 images, respectively. Similar to data splitting scheme used in BP [19], we divide the images of these three categories in the ratio of 8:2 (training vs testing) in our experiments.

HCUS [21] dataset (https://hc18.grand-challenge.org/) was from a grand challenge 2018 for fetal head segmentation. It contains 999 ultrasound images of the fetus head circumference. The corresponding ground truth are produced by a trained sonographer. The size of each 2D ultrasound image is 800×540 pixels. We divide the dataset into a training set of 800 images and a test set of 199 images in the ratio of 8:2.

4.2. Evaluation metrics and baselines

We adopt Dice score (DSC), Intersection-over-Union (IoU), and Area under curve (AUC) to measure segmentation accuracy, while Frames Per Second (FPS) and model size are used to evaluate implementing efficiency. For BP [19], we also calculate the Leaderboard Score (LB DSC) [19] of test set, provided by the Kaggle competition website, to evaluate our method. The calculation formula of LB DSC is the same as DSC, which is defined as:

$$DSC = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{5}$$

where X is the segmentation result and Y is the ground truth.

In order to show the superiority of LAEDNet, we select U-Net [11], M-Net [7], U-Net++[14], SK-U-Net [24], NAS-Unet [23], and TransUNet [15] as the baselines for comparison.

Table 2
Comparison results between state-of-the-art networks and LAEDNet on BP [19] dataset.

Model	LB DSC (%)	DSC (%)	IoU (%)	AUC (%)	p-value	Param (M)	FPS
U-Net [11]	67.0 ± 1.2	61.9 ± 0.4	57.3 ± 0.7	84.3 ± 1.7	0.001	28.97	30.7
M-Net [7]	53.9 ± 1.1	68.6 ± 0.8	64.7 ± 0.9	84.6 ± 0.4	<0.001	29.23	26.8
U-Net++ [14]	67.5 ± 0.3	70.3 ± 0.5	66.4 ± 1.2	83.9 ± 0.4	<0.001	36.15	21.5
TransUNet [15]	62.1 ± 0.3	55.1 ± 1.4	50.4 ± 1.6	82.6 ± 0.5	<0.001	419.53	4.5
NAS-Unet [23]	65.7 ± 1.0	71.8 ± 0.6	68.4 ± 0.5	74.7 ± 0.7	<0.001	**2.29**	52.6
LAEDNet-S	68.1 ± 0.2	72.4 ± 0.6	68.4 ± 1.1	88.3 ± 0.8	<0.001	4.03	**60.3**
LAEDNet-M	**69.8 ± 0.4**	73.0 ± 0.3	68.9 ± 0.2	88.6 ± 0.7	–	10.75	40.7
LAEDNet-L	69.2 ± 0.2	**73.7 ± 0.5**	**69.8 ± 0.4**	**89.3 ± 0.6**	0.017	63.25	17.0

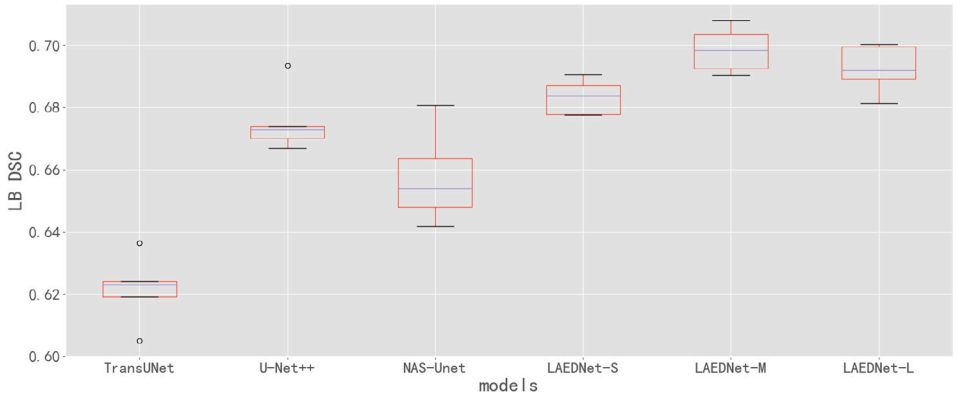

Fig. 4. Results of statistical experiments for U-Net++[14], TransUNet [15], NAS-Unet [23], LAEDNet-S, LAEDNet-M and LAEDNet-L on BP [19] dataset.

4.3. Implementation

The models are implemented using Keras framework and trained with a NVIDIA GeForce GTX 1080 GPU. The input images are all resized to 320 × 320 pixels for three datasets. The Adam optimizer [50] is used with a learning rate of 1×10^{-4}. For fair comparison, statistical tests are conducted to evaluate our model. Specifically, we randomly exchange images between training and validation set, but holding data splitting criteria in each dataset. Then our LAEDNet is retrained repeatedly in $M = 5$ times, and the average mean and standard deviation of all metrics are reported. All models are trained for 200 epochs. Moreover, we apply some data augmentation methods to improve performance, including horizontal flip and affine transformation. Since BP [19] contains a large number of images without neuron target, the distribution of positive and negative samples is imbalanced. To address this issue we adopt a periodic under-sampling (PUS) method to train LAEDNet, where half of images without targets are randomly selected, together with the whole images with targets, in each training epoch. In our experiments, the combined Dice loss [51] and Focal loss [52] is utilized to training our model, which is defined as:

$$L = L_D + L_F \tag{6}$$

4.4. Evaluation results

In this section, we demonstrate our network on BP [19], BUSI [20], and HCUS [21]. In addition, we compare the results of our LAEDNet with some state-of-the-art networks.

(1) Results on BP: Table 2 reports the comparison results between state-of-the-art networks and LAEDNet on BP [19] dataset. Our LAEDNet-M model achieves best trade-off between segmentation results and running speed. It only has 10.75M model parameters, yet obtains 69.8% LB DSC, 73.0% DSC, 68.9% IoU, and 88.6% AUC, respectively, with 40.7 FPS inference speed. LAEDNet-S is nearly 2.5 times smaller and 1.5 times faster with respect to LAEDNet-M. The segmentation accuracy, however, is only 68.1%, 72.4%, 68.4%, and 88.3% in terms of LB DSC, DSC, IoU, and AUC, respectively. Note LAEDNet-L is with 63.25M model size and only 17 FPS implementing speed, together with slightly performance drops in terms of LB DSC (0.6%). Among the baselines, TransUNet [15] is ranked at bottom in terms of segmentation accuracy, probably due to the shortage of training data. On the other hand, due to the powerful network architecture, U-Net++[14] achieves best performance in terms of LB DSC, DSC, IoU, and AUC (67.5%, 70.3%, 66.4%, and 83.9%), respectively, but is with the sacrifice of larger model size (36.15M) and lowest running speed (21.5 FPS). Although NAS-Unet [23] has the smallest model size, yet it delivers poor segmentation accuracy than LAEDNet-S (2.6%, 0.6% and 13.6% in terms of LB DSC, DSC and AUC, respectively).

Fig. 4 plots the boxplot results on BP [19], where U-Net++,[14] TransUNet [15] and NAS-Unet [23] are selected as baselines. As shown in Fig. 4, our method, especially LAEDNet-M, achieves highest segmentation accuracy. Among all baselines, TransUNet [15]

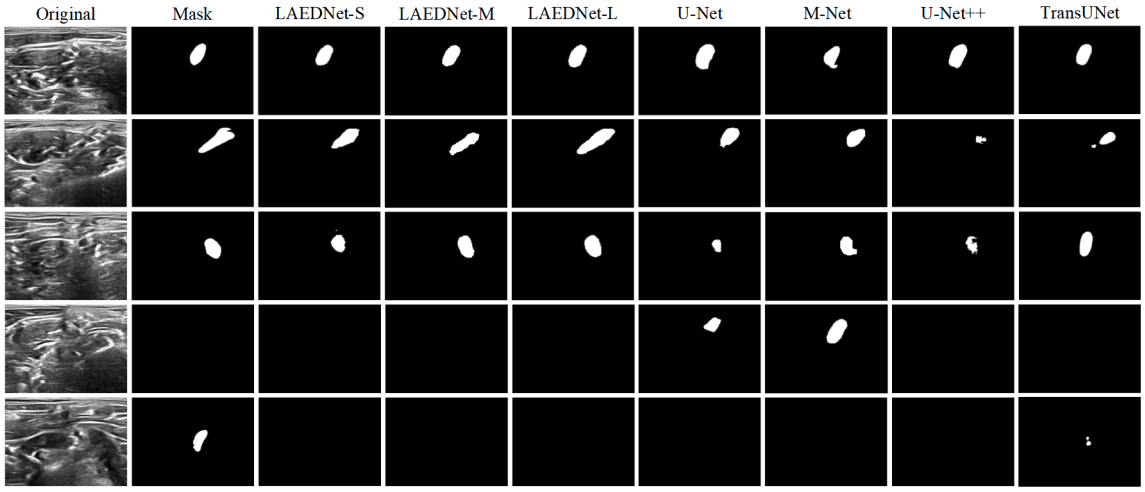

Fig. 5. Some visual examples of qualitative segmentation outputs on BP [19] dataset. From left to right are results of U-Net [11], M-Net [13], U-Net++[14], TransUNet [15], LAEDNet-S, LAEDNet-M and LAEDNet-L. The last row shows an example of poor segmentation prediction.

Table 3
Comparison results between state-of-the-art networks and LAEDNet on BUSI [20] dataset.

Model	DSC (%)	IoU (%)	AUC (%)	*p*-value	Param (M)	FPS
U-Net [11]	63.2 ± 1.1	54.2 ± 3.9	84.5 ± 2.0	<0.001	28.97	30.7
M-Net [7]	53.9 ± 2.5	45.5 ± 3.7	84.0 ± 1.6	<0.001	29.23	26.8
U-Net++ [14]	70.9 ± 1.8	63.8 ± 2.7	90.4 ± 1.3	0.201	36.15	21.5
TransUNet [15]	56.4 ± 3.3	46.6 ± 2.8	82.3 ± 1.2	<0.001	419.53	4.5
SK-U-Net [24]	72.4 ± 2.4	65.5 ± 4.4	86.4 ± 2.5	0.394	**3.94**	19.6
LAEDNet-S	73.7 ± 2.5	65.8 ± 3.6	90.0 ± 1.2	0.687	4.03	**60.3**
LAEDNet-M	73.8 ± 1.3	65.8 ± 2.0	91.0 ± 0.5	–	10.75	40.7
LAEDNet-L	**75.0 ± 0.9**	**67.6 ± 3.4**	**91.3 ± 1.1**	0.353	63.25	17.0

ranks at bottom, only obtaining 62.1% LB DSC. Additionally, NAS-Unet [23] has greatest standard deviation, indicating that it is vulnerable to fluctuations of training data.

Fig. 5 demonstrates the prediction results of U-Net [11], M-Net [13], U-Net++[14], TransUNet [15], LAEDNet-S, LAEDNet-M and LAEDNet-L on BP [19]. It can be seen that our models can predict the shape of the target most similar to the ground truth, and the contour edge is smoother. In comparison, the contour edges of the target predicted by U-Net [11], M-Net [13], U-Net++[14] and TransUNet [15] are irregular. In addition, from the example in the penultimate row in Fig. 5, some selected baselines (U-Net [11] and M-Net [13]) show relatively poor predictions for images that do not contain targets, while our method correctly classifies all pixels into backgrounds. In last row of Fig. 5, we show one example of poor segmentation prediction. This is probably because the spatial texture of target is not easily distinguish from background that has many noise and spots.

(2) Results on BUSI: The quantitative results on BUSI [20] are reported in Table 3. As can be seen, LAEDNet-S outperforms the U-Net++[14] by 2.8% and 2.0% in DSC and IoU, respectively, while the AUC is slightly lower than U-Net++[14]. LAEDNet-M outperforms the U-Net++[14] by 2.9%, 2.0%, and 0.6% in DSC, IoU, and AUC, respectively. And LAEDNet-L outperforms the U-Net++[14] by 4.1%, 3.8%, and 0.9% in DSC, IoU, and AUC, respectively. In spite of having smaller amount of parameters, SK-U-Net [24] performs slower (only 19.6 FPS) due to its fragmented operators. Besides, it obtains lower DSC, IOU and AUC scores than our LAEDNet. Specifically, LAEDNet improves the SK-U-Net [24] by a large margin of 1.3%, 0.3% and 3.6%, respectively, in terms of DSC, IOU and AUC. Overall, LAEDNet-M achieves best trade-off between accuracy and efficiency. Note that since all input images are resized into 320 × 320, we obtain the same model parameters and FPS on BP [19] and BUSI [20].

Fig. 6 shows the qualitative results of some visual examples on BUSI [20]. The results demonstrate that our LAEDNet is able to produce more fine segmentation results. The last row also exhibits a visual example of poor segmentation estimation. This is probably because that due to adhering to surrounding tissues, malignant tumors have irregular shapes and blurred boundaries, making them hard to be segmented. Even so, compared with selected baselines, the estimated outputs of LAEDNet are closer to the ground truth.

(3) Results on HCUS: Table 4 shows the results on HCUS [21], and compares with selected baselines. As can be seen, our method still achieves best performance on HCUS [21]. In particular, compared with U-Net++[14], LAEDNet-S achieves 0.4%, 0.3%, and 0.4% improvement in DSC, IoU, and AUC respectively. LAEDNet-M improves the results by 2.4%, 3.0%, and 0.6% in DSC, IoU, and AUC, respectively. LAEDNet-L can achieve 91.7% DSC, 86.1% IoU, and 99.1% AUC, which outperforms previous networks by a large margin. The model parameters and the FPS are the same as those on BP [19] and BUSI [20]. The results demonstrate

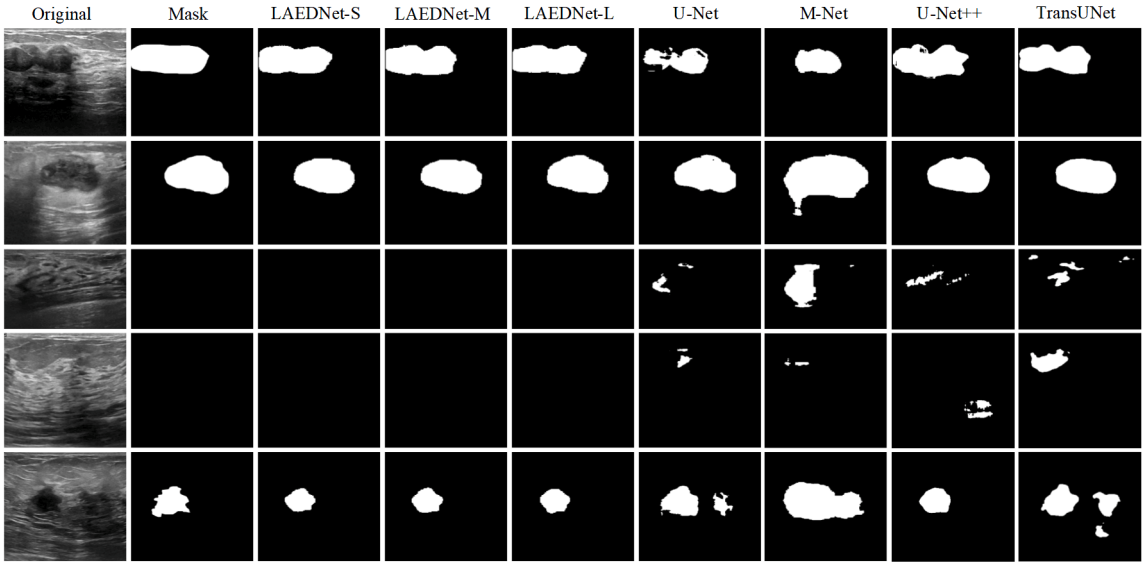

Fig. 6. Some visual examples of qualitative segmentation outputs on BUSI [20] dataset. From left to right are results of U-Net [11], M-Net [13], U-Net++[14], TransUNet [15], LAEDNet-S, LAEDNet-M and LAEDNet-L. The last row shows an example of poor segmentation prediction.

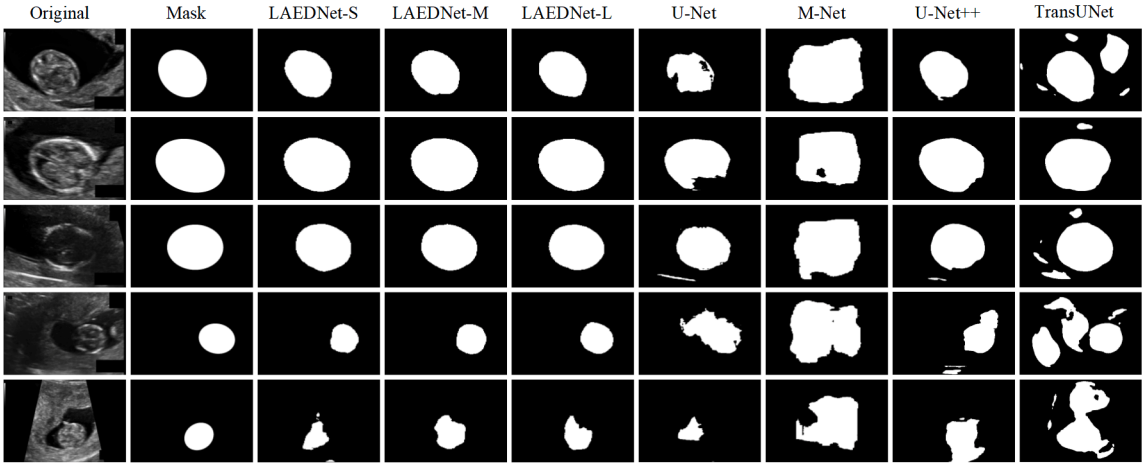

Fig. 7. Some visual examples of qualitative segmentation outputs on HCUS [21] dataset. From left to right are results of U-Net [11], M-Net [13], U-Net++[14], TransUNet [15], LAEDNet-S, LAEDNet-M and LAEDNet-L. The last row shows an example of poor segmentation prediction.

Table 4
Comparison results between state-of-the-art networks and LAEDNet on HCUS [21] dataset.

Model	DSC (%)	IoU (%)	AUC (%)	*p*-value	Param (M)	FPS
U-Net [11]	82.8 ± 0.9	73.8 ± 2.0	96.8 ± 0.8	<0.001	28.97	30.7
M-Net [7]	68.5 ± 2.6	55.5 ± 2.2	93.3 ± 1.5	<0.001	29.23	26.8
U-Net++ [14]	88.9 ± 1.8	82.6 ± 2.1	98.4 ± 0.8	0.106	36.15	21.5
TransUNet [15]	75.8 ± 3.2	64.8 ± 2.6	91.0 ± 2.4	<0.001	419.53	4.5
LAEDNet-S	89.3 ± 2.7	82.9 ± 3.5	98.8 ± 0.6	0.317	**4.03**	**60.3**
LAEDNet-M	91.3 ± 2.2	85.6 ± 2.4	99.0 ± 0.3	–	10.75	40.7
LAEDNet-L	**91.7 ± 1.5**	**86.1 ± 2.5**	**99.1 ± 0.4**	0.722	63.25	17.0

that our networks not only achieve significant improvements in segmentation accuracy, but also lead to great reduction of network parameters. Several visual examples of segmentation outputs are shown in Fig. 7. Compared with previous models, our networks show outstanding segmentation capabilities on HCUS [21].

Table 5
The results of ablation experiments.

Model	SE	PUS	LB DSC (%)	DSC (%)	IoU (%)	AUC (%)	*p*-value	Param (M)
LAEDNet-M			67.4 ± 1.2	71.8 ± 0.9	68.0 ± 1.0	84.8 ± 0.6	0.005	10.45
LAEDNet-M	✓		68.2 ± 0.6	72.2 ± 0.8	68.2 ± 0.7	86.1 ± 1.0	0.006	10.75
LAEDNet-M	✓	✓	**69.8 ± 0.6**	**73.0 ± 0.4**	**68.9 ± 0.4**	**88.6 ± 0.4**	–	10.75

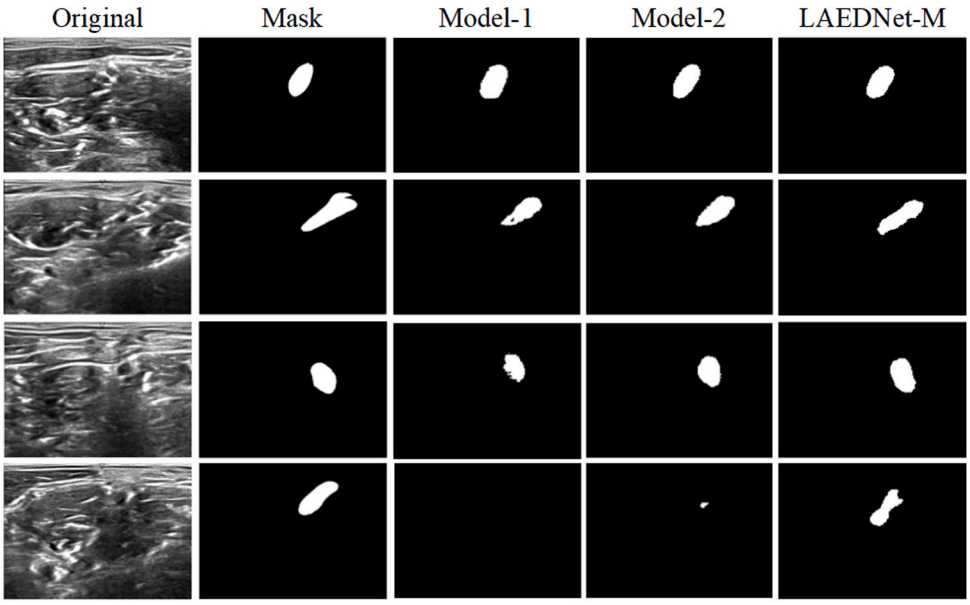

Fig. 8. Sample test predictions of Model-1(without SE module or PUS), Model-2(With SE module but without PUS) and LAEDNet-M on BP [19]. The last row is the worst segmentation results on BP [19].

5. Discussion

Although the focus of this paper is to segment target region in ultrasound image, the segmentation results are truly useful for clinical diagnose. As shown in the experimental results, LAEDNet can segment the neurons or lesions in the ultrasound image in a timely fashion. This is very helpful to ensure the success of operations such as anesthesia injection or ultrasound puncture by the doctor. In particular, in the detection of masses, such as the detection of breast nodules, LAEDNet can segment objects with clear boundaries from low-contrast images, so that doctors can judge whether it is benign or malignant according to the shape of the segmented mass.

To further verify the contributions of LRSE–SE module in decoder PUS training strategy, we carry on ablation studies on BP [19] dataset using LAEDNet-M, as it can achieves best trade-off between segmentation accuracy and implementing efficiency. As we have mentioned in Section 4.1, the region area of neuron target is relatively small with respect to the entire image, and BP datset even contains a large number of images without neuron targets, thus leading to unbalanced distribution of training samples in the dataset. To resolve this problem, we resort to PUS. For fair comparison with SE module, it is also repeatedly performed 5 times along with SE introduced. The experimental results are shown in Table 5.

As we can seen from Table 5, SE only brings 0.3M increase of model size, yet it averagely yields 0.8%, 0.4%, 0.2% and 1.3% improvement of LB DSC, DSC, IoU and AUC score, respectively, demonstrating the effectiveness of SE. On the other hand, Table 5 shows that the performance boosts 1.6%, 0.8%, 0.7% and 2.5%, respectively, using PUS training strategy, indicating that PUS well addresses unbalance issue of training data. The qualitative results are shown in Fig. 8, where Model-1 represents the baseline without SE module and PUS, and model-2 represents the model with SE but without PUS. As shown in Fig. 8, compared with baseline, employing SE produces better segmentation outputs. When PUS training strategy is introduced, our LAEDNet-M is able to yield more consistent segmentation results with ground truth.

Finally, to show whether the differences between algorithms are statistically significant, we have also conducted t-test on the results and reported p-values from Table 2 to Table 4. More specifically, for all datasets, the *p*-value is computed from each segmentation network with respect to LAEDNet-M model. It is worth to mention that, for BP dataset, we report *p*-value from the metric of LB DSC, as it is used to evaluate segmentation results on test set. While for other two datsets, we report averaged *p*-value from the metrics of DSC, IoU and AUC, respectively, as these metrics are used on validation set.

From Table 2, it is evident that, except LAEDNet-L, our LAEDNet-M model is significantly different from selected baselines. We also observe similar results in ablation study, as shown in Table 5. However, as shown in Tables 3 and 4, it is interesting that the most

similar models with respect to LAEDNet-M are LAEDNet-S and LAEDNet-L on BUSI and HCUS datasets, respectively. Furthermore, the p-value of HCUS dataset is 0.722, indicating that our LAEDNet-M model achieves similar performance with high accuracy model, yet with fewer model size and higher running speed.

6. Conclusion

In this paper, we have presented a novel asymmetrical encoder–decoder architecture, LAEDNet, to achieve best available trade-off between segmentation accuracy and implementing efficiency for ultrasound image segmentation. We have adopted 3 versions of EfficientNet as encoder to construct 3 versions of LAEDNet. A set of LRSEs are designed to couple with encode, where channel-wised attention is adopted to integrate decoding features and counterpart features in encoder. To evaluate our models, the experiments are conducted on three challenging datasets: BP [19], BUSI [20] and HCUS [21]. The experimental results demonstrate that our LAEDNet outperforms U-Net [11], M-Net [13], U-Net++[14], TransUNet [15], and other lightweight networks [23,24], in terms of segmentation accuracy. Besides, our LAEDNet-M achieves best trade-off in terms of segmentation accuracy and implementing efficiency on three datasets. We have also proved the validity of PUS training strategy and the SE module in LRSE through ablation studies. In the future, we will explore the possibility of more lightweight encoder instead of using EfficientNet. We are also interesting in applying our model to video sequence, which facilitate to online medical diagnose.

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.

Acknowledgments

This work was jointly supported in part by the National Natural Science Foundation of China under Grants 61876093, 61871234, China Postdoctoral Science Foundation Funded Project under Grant 2020M671595, and Post-doctoral Science Foundation of Jiangsu Province, China under Grant 2020Z198.

References

[1] Vashishtha V, Aju D. Nerve segmentation in ultrasound images. In: IEEE conference on innovations in power and advanced computing technologies. 2018, p. 1–5.
[2] Xue C, Zhu L, Fu H, Hu X, Li X, Zhang H, Heng P-A. Global guidance network for breast lesion segmentation in ultrasound images. Med Image Anal 2021;70:1–12.
[3] Bozidear, Potocnik, Zhauzla D. Automated analysis of a sequence of ovarian ultrasound images. Part I: segmentation of single 2D images. Image Vis Comput 2002;20(3):217–25.
[4] García H, Giraldo JJ, Álvarez M, Orozco AA, Salazar D. Peripheral nerve segmentation using speckle removal and Bayesian shape models. In: Iberian conference on pattern recognition and image analysis. 2015, p. 387–94.
[5] Lu H, Zhang M, Xu X, Li Y, Shen HT. Deep fuzzy hashing network for efficient image retrieval. IEEE Trans Fuzzy Syst 2021;29(1):166–76.
[6] Lu H, Li Y, Chen M, Kim H, Serikawa S. Brain intelligence: go beyond artificial intelligence. Mobile Netw Appl 2018;22(3):368–78.
[7] Abraham N, Illanko K, Khan N, Androutsos D. Deep learning for semantic segmentation of brachial plexus nervesin ultrasound images using U-Net and M-Net. In: IEEE international conference on imaging, signal processing and communication. 2019, p. 85–9.
[8] Vaze S, Xie W, Namburete A. Low-memory CNNs enabling real-time ultrasound segmentation towards mobile deployment. IEEE J Biomed Health Inform 2020;24(4):1059–69.
[9] Yoshiki N, Huimin L, Yujie L, Tohru K. WideSegNeXt: Semantic image segmentation using wide residual network and next dilated unit. IEEE Sens J 2021;21(10):677–93.
[10] Jiaxing S, Yujie L, Jintong C, Huimin L, Seiichi S. Image segmentation with language referring expression and comprehension. IEEE Sens J 2020;1–8.
[11] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2015, p. 234–41.
[12] Wang R, Shen H, Zhou M. Ultrasound nerve segmentation of brachial plexus based on optimized ResU-Net. In: IEEE international conference on imaging systems and techniques. 2019, p. 1–6.
[13] Fu H, Cheng J, Xu Y, Wong D, Liu J, Cao X. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans Med Imaging 2018;37(7):1597–605.
[14] Zhou Z, Siddiquee M, Tajbakhsh N, Liang J. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 2020;39(6):1856–67.
[15] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransUNet: Transformers make strong encoders for medical image segmentation. 2020, arXiv preprint arXiv:2102.04306.
[16] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems. 2017, p. 5998–6008.
[17] Aneja D, Li W. Real-time lip sync for live 2d animation. 2019, arXiv preprint arXiv:1910.08685.
[18] Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. 2019, p. 6105–14.
[19] Kaggle competitions, ultrasound nerve segmentation. 2016, [Online]. Available: https://www.kaggle.com/c/ultrasound-nerve-segmentation.
[20] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief 2019;28:1–5, [Online]. Available: https://scholar.cu.edu.eg/?q=afahmy/pages/dataset.
[21] Van D, Dagmar DB, De K, Van GB, Carlos R. Automated measurement of fetal head circumference using 2D ultrasound images. PLoS One 2018;13(8):1–20, [Online]. Available: https://hc18.grand-challenge.org/.
[22] Amiri M, Brooks R, Rivaz H. Fine-tuning U-Net for ultrasound image segmentation: Different layers, different outcomes. IEEE Trans Ultrason Ferroelectr Freq Control 2020;67(12):2510–8.

[23] Weng Y, Zhou T, Li Y, Qiu X. Nas-unet: neural architecture search for medical image segmentation. IEEE Access 2019;7:44247–57.
[24] Byra M, Jarosik P, Szubert A, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M. Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. Biomed Signal Process Control 2020;61:102027.
[25] Wu Z, Hai J, Zhang L, Chen J, Cheng G, Yan B. Cascaded fully convolutional DenseNet for automatic kidney segmentation in ultrasound images. In: IEEE international conference on artificial intelligence and big data. 2019, p. 384–8.
[26] Wu L, Xin Y, Li S, Wang T, Heng P-A, Ni D. Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation. In: IEEE international symposium on biomedical imaging. 2017, p. 663–6.
[27] Li H, Fang J, Liu S, Liang X, Yang X, Mai Z, Van MT, Wang T, Chen Z, Ni D. Cr-unet: A composite network for ovary and follicle segmentation in ultrasound images. IEEE J Biomed Health Inform 2020;24(4):974–83.
[28] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[29] Yang X, Yu L, Li S, Wen H, Luo D, Bian C, Qin J, Ni D, Heng P-A. Towards automated semantic segmentation in prenatal volumetric ultrasound. IEEE Trans Med Imaging 2019;38(1):180–93.
[30] Horng M-H, Yang C-W, Sun Y-N, Yang T-H. Deepnerve: A new convolutional neural network for the localization and segmentation of the median nerve in ultrasound image sequences. Ultrasound Med Biol 2020;46(9):2439–52.
[31] Wen W, Wu CP, Wang YD, Chen YR, Li H. Learning structured sparsity in deep neural networks. In: Advances in neural information processing systems. 2016, p. 2074–82.
[32] Li C, Richard CJ. Constrained optimization based low-rank approximation of deep neural networks. In: European conference on computer vision. 2018, p. 732–47.
[33] Wu J, Leng C, Wang Y, Hu Q, Cheng J. Quantized convolutional neural networks for mobile devices. In: IEEE conference on computer vision and pattern recognition. 2016, p. 4820–8.
[34] Nan K, Liu S, Du J, Liu H. Deep model compression for mobile platforms: A survey. Tsinghua Sci Technol 2019;24(6):677–93.
[35] Liu Y, Chen K, Liu C, Qin Z, Wang J. Structured knowledge distillation for semantic segmentation. In: IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 2604–13.
[36] Dian Q, Jia-Jun B, Zhe L, Xin S, Sheng Z, Gu J-J, Zhi-Hua W, Lei W, Hui-Fen D. Efficient medical image segmentation based on knowledge distillation. IEEE Trans Med Imaging 2021;40(12):3820–31.
[37] Ho TKK, Gwak J. Utilizing knowledge distillation in deep learning for classification of chest X-ray abnormalities. IEEE Access 2020;8:160749–61.
[38] Wang Y, Zhou Q, Liu J, Xiong J, Gao G, Wu X, Latecki LJ. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In: International conference on image processing. IEEE; 2019, p. 1860–4.
[39] Zhou Q, Wang Y, Fan Y, Wu X, Zhang S, Kang B, Latecki LJ. AGLNet: Towards real-time semantic segmentation of self-driving images via attention-guided lightweight network. Appl Soft Comput 2020;96:106682.
[40] Jha D, Ali S, Tomar NK, Johansen HD, Johansen D, Rittscher J, Riegler MA, Halvorsen P. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. IEEE Access 2021;9:40496–510.
[41] Woo S, Park J, Lee J-Y, Kweon IS. Cbam: Convolutional block attention module. In: European conference on computer vision. 2018, p. 3–19.
[42] Jie H, Li S, Gang S, Albanie S. Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 2020;42(8):2011–23.
[43] Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Computer vision and pattern recognition. 2018, p. 7794–803.
[44] Li H, Cheng J-Z, Chou Y-H, Qin J, Huang S, Lei B. AttentionNet: Learning where to focus via attention mechanism for anatomical segmentation of whole breast ultrasound images. In: IEEE international symposium on biomedical imaging. 2019, p. 1078–81.
[45] Ding J, Huang Z, Shi M, Ning C. Automatic thyroid ultrasound image segmentation based on u-shaped network. In: IEEE international congress on image and signal processing. 2019, p. 1–5.
[46] Wang Y, Dou H, Hu X, Zhu L, Yang X, Xu M, Qin J, Heng P-A, Wang T, Ni D. Deep attentive features for prostate segmentation in 3d transrectal ultrasound. IEEE Trans Med Imaging 2019;38(12):2768–78.
[47] Xue C, Zhu L, Fu H, Hu X, Li X, Zhang H, Heng P-A. Global guidance network for breast lesion segmentation in ultrasound images. Med Image Anal 2021;70:101989.
[48] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-Unet: Unet-like pure transformer for medical image segmentation. 2021, arXiv preprint arXiv:2105.05537.
[49] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: IEEE/CVF conference on computer vision and pattern recognition. 2018, p. 4510–20.
[50] Kingma D, Ba J. Adam: A method for stochastic optimization. 2014, arXiv preprint arXiv:1412.6980.
[51] Milletari F, Navab N, Ahmadi S-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: IEEE international conference on 3D vision. 2016, p. 565–71.
[52] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42(2):318–27.

Quan Zhou received Ph.D. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China in 2013. Now he is an associated professor in the college of Telecommunications and Information engineering at Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition.

Qianwen Wang received B.S. degree in information engineering from Nanjing University of Posts and Telecommunications. She is now pursuing her master degree in Nanjing University of Posts and Telecommunications. Her research interests include medical image segmentation.

Yunchao Bao received B.S. degree in information engineering from Nanjing University of Posts and Telecommunications. He is now pursuing his master degree in Nanjing University of Posts and Telecommunications. His research interests include medical image segmentation.

Lingjun Kong is now served as a full professor in Jingling Institute of Technology. His research interests include computer vision and pattern recognition.

Xin Jin received Ph.D. degree from Beihang University. Now he is an associated professor in the department of computer science and technology, Beijing Electronic Science and Technology Institute. His research interests include computer vision and pattern recognition.

Weihua Ou received Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China. Now he is a full professor in the School of Big data and computer science, Guizhou Normal University. His research interests include computer vision and pattern recognition.