SGU-Net: Shape-Guided Ultralight Network for Abdominal Image Segmentation

Tao Lei, *Senior Member, IEEE*, Rui Sun, Xiaogang Du, Huazhu Fu, *Senior Member, IEEE*, Changqing Zhang, *Member, IEEE*, and Asoke K. Nandi, *Life Fellow, IEEE*

***Abstract*—Convolutional neural networks (CNNs) have achieved significant success in medical image segmentation. However, they also suffer from the requirement of a large number of parameters, leading to a difficulty of deploying CNNs to low-source hardwares, e.g., embedded systems and mobile devices. Although some compacted or small memory-hungry models have been reported, most of them may cause degradation in segmentation accuracy. To address this issue, we propose a shape-guided ultralight network (SGU-Net) with extremely low computational costs. The proposed SGU-Net includes two main contributions: it first presents an ultralight convolution that is able to implement double separable convolutions simultaneously, i.e., asymmetric convolution and depthwise separable convolution. The proposed ultralight convolution not only effectively reduces the number of parameters but also enhances the robustness of SGU-Net. Secondly, our SGU-Net employs an additional adversarial shape-constraint to let the network learn shape representation of targets, which can significantly improve the segmentation accuracy for abdomen medical images using self-supervision. The SGU-Net is extensively tested on four public benchmark datasets, LiTS, CHAOS, NIH-TCIA and 3Dircbdb. Experimental results show that SGU-Net achieves higher segmentation accuracy using lower memory costs, and outperforms state-of-the-art networks. Moreover, we apply our ultralight convolution into a 3D volume segmentation network, which obtains a comparable performance with fewer parameters and memory usage.**

***Index Terms*—Medical image segmentation, deep learning, ultralight convolution, adversarial shape-constraint.**

I. Introduction

MEDICAL image segmentation aims to make anatomical or pathological structures changes clearer in images and it often plays a key role in computer-aided diagnosis and smart medicine due to the great improvement in diagnostic efficiency and accuracy. To help clinicians make accurate diagnosis, it is necessary to segment some crucial organs and targets in abdomen medical images and extract features from segmented targets [1]. In particular, it is more difficult to extract discriminating features from medical images than normal RGB images since the former usually suffers from problems of blur, noise, low contrast, etc. In recent years, deep learning, especially the U-shaped encoder-decoder network [2], has been widely used in medical image segmentation due to its excellent performance. As the encoder of U-shaped network [2] used for feature learning are insensitive to image noise, blur, low contrast, etc., many improved U-shaped networks such as U-Net++ [3], mU-Net [4], Attention U-Net [5], TransUNet [6], Swin-Unet [7], etc. can provide excellent segmentation results for medical images. Although these networks gain high segmentation accuracy, they are complex due to a large number of parameters and high memory usage thus leading to the difficulty of deployment on mobile devices. How to balance the complexity of networks and segmentation accuracy is a challenge.

Fig. 1 shows the Dice value and the number of parameters of different networks on the CHAOS-CT [8] dataset. We can see that some medical image segmentation models have a huge number of parameters, e. g. R2U-Net [9] of 39.09 M, Attention U-Net [5] of 34.88 M and V-Net [10] of 65.17 M. It is clear that most of these high-accuracy networks are unsuitable to be deployed on mobile devices. Although some lightweight networks [3] have been reported, they may suffer from serious performance degradation when under the low computing resources. To address the above issues, we present a shape-guided

Manuscript received 1 May 2022; revised 27 October 2022; accepted 14 January 2023. Date of publication 19 January 2023; date of current version 7 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62271296, 61871259, and 61861024, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2021JC-47, in part by the Key Research and Development Program of Shaanxi under Grants 2022GY-436 and 2021ZDLGY08-07, in part by the Natural Science Basic Research Program of Shaanxi under Grants 2022JQ-634 and 2022JQ-018, in part by the Shaanxi Joint Laboratory of Artificial Intelligence under Grant 2020SS-03, and in part by Huazhu Fu's A*STAR Central Research Fund and AISG Tech Challenge Funding under Grant AISG2-TC-2021-003. *(Corresponding author: Xiaogang Du.)*

Tao Lei is with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China, and also with the Department of Geriatric Surgery, First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710021, China (e-mail: leitao@sust.edu.cn).

Rui Sun and Xiaogang Du are with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: siri0920@163.com; du423@sina.com).

Huazhu Fu is with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: hzfu@ieee.org).

Changqing Zhang is with the School of Computer Science and Technology, Tianjin University, Tianjin 300222, China (e-mail: zhangchangqing@tju.edu.cn).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge UB8 3PH, U.K., and also with the College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: Asoke.Nandi@brunel.ac.uk).

Digital Object Identifier 10.1109/JBHI.2023.3238183

The available code of SGU-Net is released at https://github.com/SUST-reynole/SGUNet

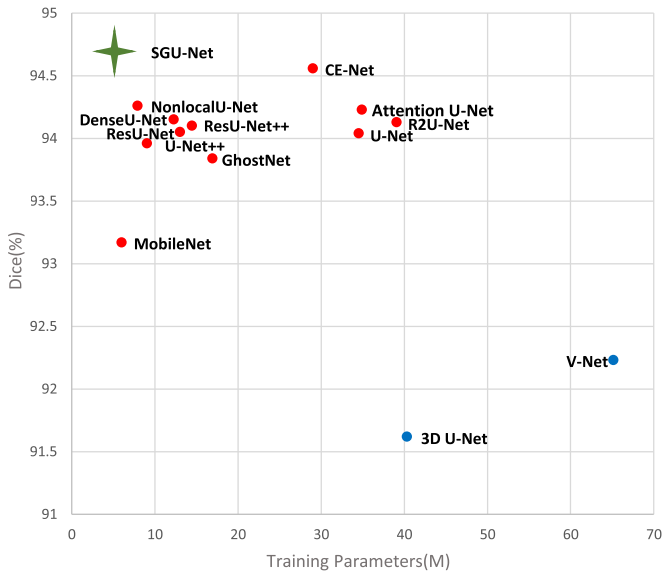

Fig. 1. The Dice value and the number of parameters of different frameworks on CHAOS-CT datasets. Blue points: 3D CNNs. Red points: 2D CNNs. Compared with methods of small memory footprint, the proposed SGU-Net locates in the left-top indicating fewer parameters. while achieving higher performance.

ultralight network (SGU-Net) with extremely low computational costs for medical image segmentation.

To improve computational efficiency and reduce the number of parameters, we propose an ultralight convolution (UC) that is a plug-and-play operation and can be used in arbitrary networks. Compared with the popular asymmetric convolution [11] and depthwise separable convolution [12], ultralight convolution has obvious advantages in reducing the number of parameters and improving feature representation ability. Specifically, for the input feature maps, the ultralight convolution first performs depthwise asymmetric convolution that consist of the cascade of $1 \times k$ and $k \times 1$ convolution. Then, the pointwise convolution is utilized to obtain the output feature maps.

To improve segmentation accurancy, we present a shape adversarial autoencoder (SAAE) that is an additional self-supervision strategy to raise the performance of our segmentation network by alternating training SAAE and the segmentation network. The proposed SAAE has a completely different and novel working mode from popular autoencoder-based shape constraint methods [13], [14]. Specifically, we try to use an autoencoder to explore the ability of CNNs on shape representation of predicted targets in low-dimensional manifold. It is worth mentioning that the proposed SAAE and segmentation network are trained cooperatively, which not only forces the proposed segmentation network to output targets with more real shape information but also is a costless supervision operation for the segmentation network.

The experimental results show that SGU-Net not only obtains higher segmentation performance but also provides better target shape prediction. Besides, the SGU-Net requires fewer parameters and lower computational costs with 4.99 M and 4.98 GFLOPs, respectively.

II. Related Work

Medical Segmentation Networks: Currently, most of medical image segmentation networks are based on U-shape architecture. These networks can be roughly grouped two categories that are often used for 2D images and 3D volumetric data, respectively.

For 2D medical image segmentation, residual and dense connections are popular for improving network performance, such as ResUNet [15], mU-Net [4], and DenseUNet [16]. The improved networks replace each submodule of U-Net in the form of residual and dense connections, respectively. This improvement can accelerate the model convergence and improve the feature reuse such as U-Net++ [3] and R2UNet [9]. It has been demonstrated that the attention mechanism is very useful for improving the feature representation ability of networks. Inspired by this, Attention U-Net [5] with spatial attention, ResUNet++ [17] with channel attention, and Non-local U-Net [18] with self-attention mechanism are proposed and used for different segmentation tasks to overcome the drawback of feature utilization in U-Net. Compared to attention mechanism, multi-scale feature fusion, for example, the atrous spatial pyramid pooling module (ASPP) [19], is also a useful way for improving network performance. By integrating ASPP into U-shape networks, both CE-Net [20] and DefED-Net [21] achieve better target segmentation in medical images.

For medical volumetric data, 2D CNNs are often limited since they ignore the temporal information of volumetric data. To overcome this drawback, 3D CNN-based models such as 3DU-Net [22] and V-net [10] have been proposed. Although these 3D networks can simultaneously explore the temporal information of inter-slice and spatial information of inner-slice, they suffer from some new problems such as more parameters, much memory usage, and much narrow reception fields than 2D networks [23].

As human organs usually have a fixed shape and position, the incorporation of the prior-knowledge about target shape and position is crucial for improving medical image segmentation effect. Mosinska et al. [24] used a pre-trained model to constrain the shape of segmentation targets. Li et al. [25] proposed a shape perception strategy based on generative adversarial networks (GANs). Lei et al. [26] proposed a network based on adversarial consistency learning and dynamic convolution. Al Arif et al. [27] used symbolic distance functions (SDFs) generated by modified U-Net instead of partition maps to obtain better topology prediction results. Furthermore, some researchers [28], [29] used autoencoder to constrain the shape of segmented targets. However, the interpretability of priori information utilization is insufficient in above methods.

Lightweight Segmentation Networks: Small medical image segmentation models require a good trade-off between segmentation accuracy and model szies for clinical mobile devices. The methods used for lightweight network design can be roughly categorized into two groups, model compression and model compacting. For a given model, the purpose of model compression is to reduce the computational costs as well as the number of parameters. Common model compression methods can be divided into three categories. The first is model pruning [30], [31],

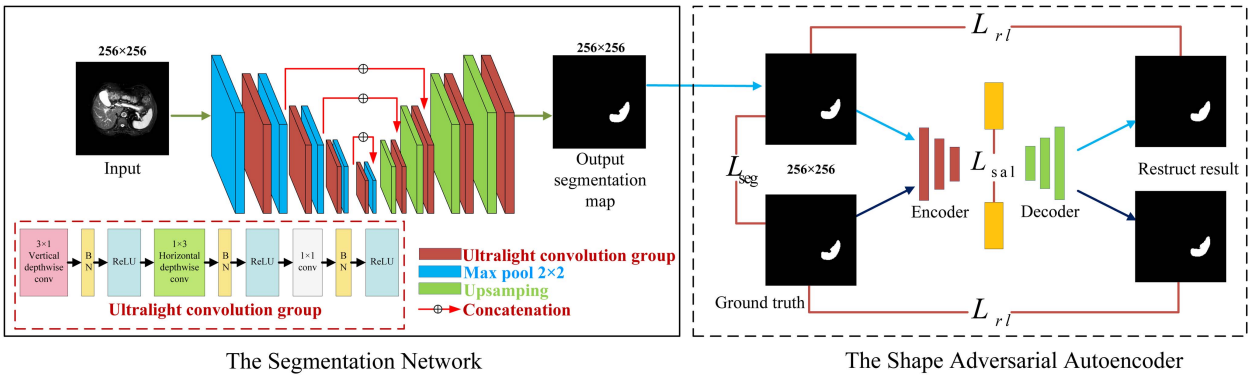

Fig. 2. The overall architecture of SGU-Net. The SGU-Net consists of two parts: the segmentation network on the left and the shape adversarial autoencoder (SAAE) on the right. SAAE encodes both segmentation results and labels into low-dimensional manifold, aiming to constrain their shape representation in low-dimensional manifold. The performance of the segmentation network is improved by adversarial training between SAAE and the segmentation network. It is worth mentioning that SAAE is an additional network that is only used in the training stage, once the training is over, the SAAE will be removed from the testing stage.

[32], and it aims to cut off unnecessary connections between different neurons for further speedup in practice. In medical image segmentation tasks, for example, U-Net++ [3] uses model pruning to reduce the number of parameters. Secondly, model quantization [33], [34] focuses on the reduction of the number of bits required on each weight to compress the original network, for example, binarization methods with only 1-bit value can greatly accelerate the model inference by efficient binary manipulation. In addition, knowledge distillation [35], [36], [37], [38] uses larger models to teach smaller models, which improves the performance of smaller models. The performance of these methods usually depends on the given pre-trained models. As for the model compaction, a lot of work has been reported in recent years. MobileNets [12], [39] proposes the depthwise separable convolution that decomposes a vanilla convolution into a depthwise convolution and a pointwise convolution. Lei et al. [40] and Zhang et al. [41] extended the depthwise separable convolution to 3D networks and applied it to medical image segmentation, effectively reducing the number of model parameters and computational costs. In addition, GhostNet [42] proposes a Ghost module to generate more feature maps from cheap operations. Lo et al. [11] and Szegedy et al. [43] used the strategy of decomposing the standard 3×3 convolution into 3×1 and 1×3 convolutions to reduce the number of parameters and computational costs at the expense of slight performance degradation. ShuffleNet [44] divides the convolutions into multiple groups in a similar way to [12], which leads to a significant reduction in FLOPs with a rather small decrease on accuracy. By combining asymmetric convolution and dilation convolution, researchers [45], [46] further designed a depthwise asymmetric dilation convolution to reduce the number of parameters of models.

Compared to previous work such as asymmetric convolution [12] and depthwise separable convolution [47], our ultralight convolution not only achieves a higher model compression ratio but also provides better feature representation ability. Compared with the current autoencoder-based shape constraint methods [13], [14], our SAAE is considered as an additional self-supervision to explore a more accurate representation of shapes in low-dimensional manifold. SAAE can provide cost-free accuracy gains since the segmentation network employed by our SGU-Net can work independently from SAAE during the testing stage.

III. Method

The overall architecture of SGU-Net is shown in Fig. 2 and that consists of two parts: the segmentation network and the shape adversarial autoencoder (SAAE). Compared to U-Net, on the one hand, the segmentation network uses the ultralight convolutional groups instead of the vanilla convolutional groups in the encoding stage. On the other hand, since the deconvolution may cause grid effect [48], which is unfavorable to pixel-level segmentation, the deconvolution in the vanilla U-Net is replaced by a combination of upsampling and ultralight convolutional groups. The shape-guide module adds additional shape constraints to the segmentation network, which encourages the predictions of segmentation network to be consistent with the shape of the organ by encoding shape information into low-dimensional manifold.

A. Ultralight Convolution

Overview: The proposed ultralight convolution tries to integrate the advantages of both asymmetric convolution and depthwise separable convolution. We factorize vanilla convolution into depthwise asymmetric convolution and pointwise convolution. For SGU-Net, depthwise asymmetric convolution applies asymmetric convolution to each input channel, and then pointwise convolution is used for channel information merging. The vanilla convolution simultaneously performs filtering on the channel and spatial dimension and merges inputs to form a new output, while the proposed ultralight convolution divides itself into three layers, namely the horizontal and vertical convolution layers for filtering and a separate layer for merging. This decomposition has a significant effect in reducing computational costs and model size.

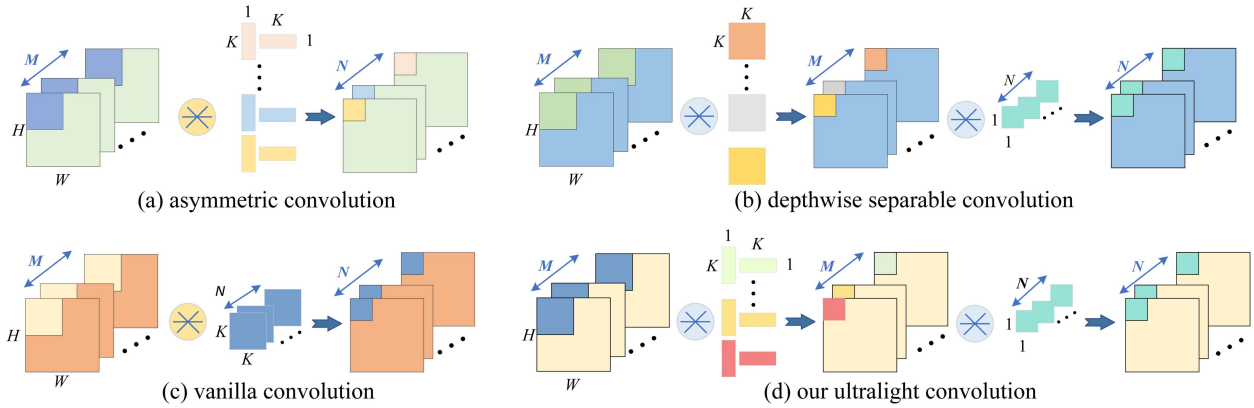

Fig. 3. The comparison between ultralight convolution and popular convolution strategies.

Fig. 3 shows the comparison between ultralight convolution and popular convolution strategies including vanilla convolution, asymmetric convolution and depthwise separable convolution. According to Fig. 3, we can see that both asymmetric convolution and depthwise separable convolution can reduce the number of parameters and computational costs compared to vanilla convolution. However, depthwise separable convolution is superior to asymmetric convolution since it achieves the decouple operation between spatial convolution and channel convolution operation leading to more lightweight networks. Compared to asymmetric convolution and depthwise separable convolution, our proposed ultralight convolution has the following advantages:

a) Compared with asymmetric convolution as shown in Fig. 3(a), our ultralight convolution includes two stages namely depthwise asymmetric convolution and pointwise convolution, it can effectively decouple the spatial and channel dimensions of convolution operation, leading to more efficient model compression.
b) Compared with depthwise separable convolution as shown in Fig. 3(b), our ultralight convolution implements depthwise asymmetric convolution while the former implements standard depthwise convolution, which achieves the spatial decomposition of convolution kernels and is especially helpful for improving feature extraction of irregular organs in abdominal image segmentation.

Complexity Analysis: The vanilla convolution takes the feature map X of size $H_x \times W_x \times M$ as input and outputs the feature map Y of size $H_y \times W_y \times N$, where H_x and W_x are the spatial height and width of X, M is the number of channels of X, H_y and W_y are the spatial height and width of Y, and N is the number of channels of Y. A vanilla convolution layer usually employs a convolutional kernels of size $D_k \times D_k \times M \times N$, where D_k is the size of the convolution kernel, M is the number of input channels, and N is the number of output channels. Consequently, the computational cost of the vanilla convolution is $D_k \times D_k \times M \times N \times H_y \times W_y$.

In fact, the vanilla convolution directly output feature maps by implementing a complex convolution operation. However, the proposed ultralight convolution factorizes a vanilla convolution into two processes. First, the input feature maps are filtered channel-by-channel using asymmetric convolution. Secondly, a pointwise convolution (1×1 convolution) is used to create a linear combination output. It is worth mentioning that both normalization and activation are also implemented in these two steps.

It is clear that the computational cost of the $k \times k$ depthwise convolution is $D_k \times D_k \times M \times H_y \times W_y$. In our ultralight convolution, we factorize the $k \times k$ depthwise convolution kernel into a $k \times 1$ kernel and a $1 \times k$ kernel. As a result, the computational cost of the $k \times k$ depthwise asymmetric convolution denoted by C_{da} is

$$C_{da} = D_k \times 1 \times M \times H_y \times W_y + 1 \times D_k \times M \times H_y \times W_y, \tag{1}$$

we can see that the further factorization of depthwise convolution can reduce the computational complexity.

As the computational cost of a pointwise convolution (1×1 convolution) is $M \times N \times H_y \times W_y$, the computational cost of the ultralight convolution denoted by C_{uc} is

$$C_{uc} = 2 \times D_k \times M \times H_y \times W_y + M \times N \times H_y \times W_y, \tag{2}$$

Although the basic ultralight convolution requires a small number of parameters and low computational costs, it may suffers from difficulties in some specific scenarios due to the requirement of smaller and faster models to mobile devices. To assure the flexibility of the proposed ultralight convolution, we introduce a hyperparameter called thinning exponent δ. The thinning exponent δ is important to compress further the model from the channel dimension and thus controls the overall size and computational efficiency of the model. When there are specific scenarios requiring smaller models and faster inference speed, we can make the models better by adjusting the thinning exponent δ. Therefore, δ is an important hyperparameter to balance model size and segmentation accuracy. In this case, the number of input channels M changes to δM and the number of output channels N changes to δN. The computational cost of an ultralight convolution with a thinning exponent δ denoted by $\hat{C}_{uc}$ is

$$\hat{C}_{uc} = 2 \times D_k \times \delta M \times H_y \times W_y + \delta M \times \delta N \times H_y \times W_y, \tag{3}$$

where $\delta \in (0, 1]$, $\delta = 1$ is the original ultralight network and $\delta < 1$ is the skinny ultralight network. The thinning exponent is a hyperparameter that can be adjusted for any model according to the desired number of parameters and segmentation accuracy.

Asymmetric convolutional validity analysis: In fact, the asymmetric convolution is often used for existing square kernel convolution layers for compression and acceleration. However, it may cause the performance degradation by factorizing the $k \times k$ convolution directly into $k \times 1$ and $1 \times k$ convolutions. One main reason is the weak extraction capability of asymmetric convolution for channel features in the case of multiple channels, as the factorization destroys the feature space extracted by the square convolution kernel leading to the loss of channel information. Unlike previous work [11], [47], our proposed ultralight convolution applies asymmetric convolution for the channel-by-channel of X, thus avoiding this drawback. Also, performing asymmetric convolution operation on each channel can better enhance the robustness of the model to prevent rotational distortions. Especially in medical images, as the shape of organ is usually irregular, the asymmetric convolution can better accommodate irregular shapes and extract more effective features for abdominal organ segmentation than a vanilla convolution.

B. Shape-Guided Strategy

Overview: For medical image segmentation, the predicted target contours are very important since these results are often used for 3D organ reconstruction. However, it is difficult to segment targets accurately due to the limitation of imaging quality. Therefore, the strategy of adding higher-level shape constraints to a segmentation network is a solution that can make prediction results more consistent with prior anatomical knowledge. Nevertheless, a ground truth usually involves structural and high-dimensional information, and measuring the shape similarity in a high-dimensional space is extremely difficult. To solve the above problems, we present SAAE helping the segmentation framework to explore the shape representation and constraints, and to guide prediction results of the segmentation network to be close to the ground truth. Specifically, SAAE is a trainable neural network to capture the salient features of the input shape and encodes them into low-dimensional manifold. If SAAE can reconstruct the input shape well, the encoding in low-dimensional manifold can be well approximated as a representation of the shape features.

Training process: The overall training stage can be seen in Fig. 2. The motivation of SAAE consists of two parts. The first is to insert an additional shape guidance strategy into the segmentation model to improve the segmentation accuracy without increasing network parameters. The second is to make full use of the rich prior knowledge of abdominal images to improve the interpretability of model learning. Our SAAE explores the representation of shape in a microscopic way, and minimizes the difference between prediction results and labels through the gradient backpropagation algorithm. This is because the shape representation of abdominal organ is high-dimensional information, it is difficult to directly measure the shape difference between the prediction results and the labels. To solve this problem, our SAAE uses an autoencoder to explore the shape representation abdominal organ in low-dimensional manifold. At the same time, since the organ labels only contain shape and position information, our SAAE can further encode the shape and position information of organs after reconstructing a large number of labels and prediction results. Moreover, the constantly trained SAAE can capture the subtle difference between the segmentation result and the real organ shapes, which can be used to monitor the segmentation network to output better results.

SAAE is trained by reconstructing the prediction results and labels from the segmentation network. It contains two loss functions. The first loss function L_{rl} learns shape representation by minimizing the difference between the reconstructed shape and the input shape, and the second loss function L_{sal} tries to distinguish the difference between the predicted shape and the real shapes by maximizing the representation of the shape in low-dimensional manifold. The two loss functions help SAAE better to encode the predicted shape and the real shape of organ, capture the subtle difference between them, and force the segmentation network to segment the results close enough to the real organ shape. The loss function of the segmentation network is expressed as L_{seg}. Therefore, the optimization objective of the overall segmentation network G and shape adversarial autoencoder D is

$$\boldsymbol{T} = \min_{\mathbf{G}} \max_{\mathbf{D}} (L_{seg} + L_{rl} + \theta L_{sal}). \tag{4}$$

In fact, G and D are done by alternating training, and they are like playing a game against each other. First, G is optimized by fixing D and minimizing subsequent losses

$$\boldsymbol{T}_1 = \min_{\mathbf{G}} (L_{seg} + \theta_1 L_{sal}), \tag{5}$$

G is encouraged to segment the image closer to the real label by optimizing G. Then D is optimized by fixing G

$$\boldsymbol{T}_2 = \max_{\mathbf{D}} (-L_{rl} + \theta_2 L_{sal}). \tag{6}$$

For our model training, G and D are implemented alternate training. The segmentation network and SAAE is like playing a minimum-maximum game. In low-dimensional manifold, the segmentation network tries to get a result consistent with the real shape to minimize the distance between output results and labels, while SAAE tries to learn better encoding and feature extraction methods to maximize the distance between output results and labels. In other words, our SAAE needs to maximize its ability to find the difference between labels and segmentation results, while the segmentation network tries to cheat SAAE by minimizing the difference. During the whole training process, the organ contours predicted by the segmentation network G will continuously approach the real organ contours. When our SAAE cannot distinguish the difference between these contours, D no longer provides effective supervision. At this time, the segmentation network G can output better segmentation results independently. Our SAAE thus has two advantages. First, it can represent shapes in different ways and uses the gradient

backpropagation algorithm to optimize the shape-guided segmentation network. Second, our SAAE can distinguish the subtle difference among different shapes, so that the shape predicted by the segmentation network can be closer to the real organ shape.

SAAE validity analysis: Previous work [24], [49] on shape constraints typically use a pre-trained model to guide shape constraints or a classification discriminator to distinguish between true and false labels. However, the parameters of the pre-trained model are fixed and the model cannot serve to discriminate when the predicted contours are close to the real shapes. Besides, the classification discriminator cannot potentially encode the contours and does not guide the segmentation network well enough to constrain it by true or false signals. The SAAE is able to represent contour information in a microscopic way and can regularize the estimated target contours by minimizing the difference between the predicted result and the ground truth. In addition, the trained SAAE is able to distinguish subtle difference between contours, and it still gives correct penalties even if the result predicted by the segmentation network is closer to the ground truth shape.

C. Loss Function

In our task, there are three loss functions that are segmentation loss L_{seg}, reconstruction loss L_{rl} and shape adversarial loss L_{sal}. First, the standard cross entropy loss L_{cross} and the boundary loss L_{bd} [50] are used in L_{seg}. However, since the boundary loss is unstable and easily leads to training difficulties, the final loss function of segmentation network is defined as

$$L_{seg} = L_{cross} + \alpha L_{bd}. \tag{7}$$

We define the reconstruction loss of SAAE as

$$L_{rl} = (L_{cross} + \alpha L_{bd})_{y}^{D(y)} + (L_{cross} + \alpha L_{bd})_{G(x)}^{D(G(x))}, \tag{8}$$

where x is the input image, y is its corresponding ground truth, G is the segmentation network, $G(x)$ is the segmentation result corresponding to x, $D(y)$ and $D(G(x))$ are the reconstruction results of SAAE corresponding to the ground truth y and the predicted result $G(x)$.

For shape adversarial loss, since it is a shape representation in low-dimensional manifold, we define L_{sal} as

$$L_{sal} = \sum_{i=1}^{n} (E(y) - E(G(x)))^2, \tag{9}$$

where $E(\cdot)$ is the encoding of the predicted shape of the segmentation network with the shape of ground truth.

IV. Experiments

A. Datasets and Pre-Processing

In our experiments, the Combined (CT-MR) healthy abdominal organ segmentation (CHAOS) [8] and the Liver Tumor Segmentation Challenge (LiTS) [51] are considered as experimental datasets. The CHAOS from the CHAOS challenge is collected by the Department of Radiology, Dokuz Eylul University Hospital, Izmir, Turkey. It contains a total of 80 cases, in which 40 cases are abdominal CT scans containing ground truth of liver segmentation, and the other 40 cases are T1-DUAL in phase (T1-DUALin). Three radiologists (10, 12 and 28 years of experience) are involved in the manual segmentation. The final masks are obtained by using majority vote, which ensures the accuracy of the ground truth. We divided the CT and MR images into training set, validation set and testing set in a ratio of 6:2:2, respectively. The MR images are 256×256 or 288×288 in size with axial slice numbers ranging from 26 to 50 and layer thicknesses between 4.4 and 8.0 mm, and the CT images are 512×512 in size with axial slice number ranging from 78 to 294 and layer thickness between 2.0 mm and 3.2 mm. Training data are subjected to random scaling, rotation, cropping and shifting operations. In our experiments, the given models are dedicated to a single modality (T2-SPIR, CT) and a single organ (liver, right kidney, left kidney, spleen). Thus, each model performs binary rather than multiclass segmentation to extract robust organ-specific features.

The LiTS includes 131 labeled 3D CT scans, where the resolution in-plane ranges from 0.55 mm to 1.0 mm and slice spacing ranges from 0.45 mm to 6.0 mm. We constructed the training set and validation set using 90 patients (total 43,219 images) and 10 patients (total 1,500 images), respectively. Then the other 30 patients (total 15,419 images) are considered as the testing set. It is worth mentioning that the LiTS dataset does not use data augmentation techniques.

Medical CT images are different from natural images, the former is able to obtain wider range of values from -1000 to 3000 than the latter from 0 to 255. To remove interferences and enhance liver areas, we truncated the image intensity values of all scans of $[-200, 250]$ HU.

B. Experimental Setup and Evaluation Metrics

All models are trained by the framework of Pytorch 1.3.0 and implemented on a desktop PC with double NVIDIA GeForce RTX 2080 Ti with 11 GB RAM. The initial learning rate (lr) is set to 0.001, and then decays according to the poly schedule $lr = lr \times (1 - iterations/totaliterations)^{0.9}$. We used the Adam gradient descent with momentum to optimize all models.

The hyperparameters in SGU-Net are set as follows: the thinning exponent δ is set to 0.25 since we aim to obtain an ultralight network as soon as possible, where $0 < \delta \leq 1$. If the value of δ is too large, then the model size will be also large. We presented more details on the set of the value of δ in the section of our discussion. θ_1 in (5) is set to 5, θ_2 in (6) is set to 0.01 and α in (7) is set to 0.5.

We denoted the segmentation result by S and the ground truth by G. Dice value is estimated by $2(S \cap G)/(|S| + |G|)$, where the Dice value of in the interval [0,1]. A perfect segmentation yields the Dice value is 1. In addition, average/maximum symmetric surface distances (ASSD/MSSD) [51] corresponds to the average/maximum Hausdorff distance between border voxels in S and G. Dice generates an overlap measure while ASSD and MSSD are surface distance measures. The former focuses more on the interior of segmentation targets, while the latter focuses more on the shape similarity of segmentation targets. It is worth noticing that we calculated the above metrics in the binary slice segmentation results.

TABLE I
COMPARISON OF ABLATION EXPERIMENTS OF UC AND SAAE ON THE CHAOS-CT TESTING SET

Method	ASSD (mm)↓	MSSD (mm)↓	Dice(%)↑	Para. (M)↓
U-Net [2]	1.70±0.94	29.52±12.23	94.04±2.32	34.53
U-Net + AC [43]	1.71±0.83	28.75±12.35	93.86±2.08	15.67
U-Net + DC [12]	1.72±0.60	28.64±12.13	93.17±2.46	5.99
U-Net + UC	1.67±0.92	28.60±12.40	94.12±1.99	**4.99**
U-Net + GAN	1.61±0.75	28.12±12.06	94.41±2.68	34.53
U-Net + AE	1.63±0.78	28.20±13.15	94.38±2.74	34.53
U-Net + SAAE	1.57±0.48	27.02±14.50	94.56±1.75	34.53
U-Net + DC + SAAE	1.60±0.57	27.48±13.26	94.32±1.84	5.99
SGU-Net	**1.43±0.35**	**26.26±11.9**	**94.68±1.64**	**4.99**

The best values are in bold.

C. Ablation Studies

In this paper, two contributions are highlighted, one is the replacement of vanilla convolution by ultralight convolution, and the other is the design of a shape adversarial autoencoder to impose additional shape constraints on the segmentation network. To demonstrate the effectiveness of our contributions, we performed ablation experiments on the CHAOS liver dataset and the NIH-TCIA pancreas dataset.

The results in Table I demonstrate the validity of our contributions. For simplicity, AC represents asymmetric convolution [43], and DC denotes depthwise separable convolution [12]. From the experimental results, compared with the vanilla U-Net, U-Net+AC and U-Net+DC, the U-Net+UC can significantly reduce the number of parameters while improving the Dice value. The vanilla U-Net achieves mean Dice value of 94.04%, ASSD value of 1.70 mm, and MSSD value of 29.52 mm. After replacing the vanilla convolution with ultralight convolution, the parameters of U-Net decrease by 29.54 M, the Dice value increases by 0.08%, the value of ASSD and MSSD decrease by 0.03 mm and 1.46 mm, respectively. This shows that our ultralight convolution not only effectively reduces the number of parameters, but also improves the segmentation accuracy due to the fact that the asymmetric convolution can better adapt to irregular organ shapes in abdominal images.

To show the validity of SAAE, we compared U-Net with U-Net+SAAE, and compared U-Net+UC with U-Net+UC+SAAE. We can see that the Dice value increases by 0.52% and 0.56%, ASSD value decreases by 0.13 mm and 0.24 mm, and MSSD value decreases by 2.5 mm and 2.34 mm, respectively. By using SAAE with U-Net, it not only provides better shape guidance, but also significantly reduces the ASSD and MSSD values.

Meanwhile, we also compared the results of SAAE with the standard generative adversal network (GAN) and autoencoder (AE) actions in our ablation experiments of Table I, respectively. Compared to the GAN constraint, the results provided by SAAE are 0.04 mm and 1.1 mm lower in ASSD and MSSD, respectively. This shows that using only the binary signal of the discriminator in the GAN does not provide a good constraint. Similarly, compared to the use of AE, the ASSD and MSSD of the results provided by SAAE are reduced by 0.06 mm and 1.18 mm, respectively. This demonstrates that the encoding of labels and segmentation results in SAAE is important, and a well-trained autoencoder can well encode the differences in labels and segmentation results into the low-dimensional space and use the difference L_{sal} in Fig. 2 between them to supervise the segmentation network.

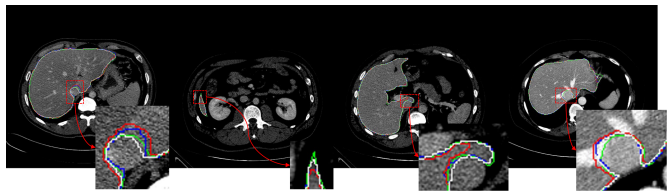

Fig. 4. Comparison of segmentation boundaries in ablation studies. The green, red, blue and grey denote ground truth, the result provided by U-Net, the result provided by U-Net+SAAE and the result provided by SGU-Net, respectively.

TABLE II
ABLATION EXPERIMENTS OF SAAE ON THE CHAOS-CT TESTING SET USING LIGHTWEIGHT U-NET

Method	ASSD (mm)↓	MSSD (mm)↓	Dice(%)↑	Para. (M)↓
U-Net	1.75±0.83	31.26±13.41	92.36±2.87	8.56
U-Net + SAAE	**1.72±0.81**	**30.59±12.22**	**93.12±2.69**	**8.56**

The best values are in bold.

TABLE III
ABLATION EXPERIMENTS OF UC AND SAAE ON THE NIH-TCIA CT TESTING SET (U-NET IS THE BACKBONE NETWORK)

Method	RMSE (mm)↓	JA (mm)↑	Dice(%)↑	Para. (M)↓
U-Net	6.15±5.32	66.42±9.45	79.33±7.24	34.53
U-Net + UC	5.66±5.17	66.98±9.21	79.54±7.01	4.99
U-Net + UC + SAAE	**4.12±3.71**	**69.07±8.62**	**80.72±6.53**	**4.99**

The best values are in bold.

Fig. 4 shows the comparison of segmentation boundaries, which further illustrates the ablation studies. As can be seen from the results, the utilization of SAAE supervision can effectively provide shape supervision and constraint without extra parameters and computational costs. And the segmentation results are further improved with the help of ultralight convolution.

To further verify the performance of SAAE, we performed the lightweight U-Net network with four stages on the CHAOS dataset. The experimental results are shown in Table II. We find that our SAAE can improve segmentation accuracy by 1.24% (Dice value). Therefore, our SAAE can obtain a clear improvement in segmentation accuracy for low Dice segmentation tasks. It is a general module that can be combined with different backbone networks for different segmentation tasks.

In addition, we conducted an additional experiment on a pancreatic dataset with a low Dice value to demonstrate further our contributions. The NIH-TCIA CT dataset [52] comes from the US National Institute of Health (NIH), which contains 82 abdominal enhanced 3D CT scans. In the direction of the axial viewpoint, the CT slice size is 512×512 pixels, and the number of slices varies from 181 to 466 for different patients. The experimental results are shown in Table III.

TABLE IV
SEGMENTATION RESULTS OBTAINED BY SGU-NET WITH DIFFERENT HYPERPARAMETERS ON THE CHAOS-CT DATASET

Model	θ_1	θ_2	α	Dice(%)
SGU-Net	1	0.1	1	94.02±2.96
	3	0.05	1	94.17±2.42
	5	0.01	1	94.42±1.87
	1	0.1	0.5	94.10±3.06
	3	0.05	0.5	94.23±2.85
	5	**0.01**	**0.5**	**94.68±1.64**

The best values are in bold.

This experiment has two purposes. On the one hand, we aim to show that our ultralight convolution module is effective and universal. On the other hand, we want to demonstrate that our SAAE module is effective for abdominal organ segmentation whether organs are large or small. It can be seen in Table III that ultralight convolution has a good universality, which can effectively extract target features and plays an outstanding role. It greatly reduces the complexity of the model and improves the segmentation accuracy. For relatively fixed abdominal organs, SAAE can achieve shape guidance and make full use of prior knowledge to guide the segmentation network to obtain better abdominal organ segmentation results.

In addition to conducting ablation experiments on our contributions UC and SAAE, we also conducted additional ablation experiments on three hyperparameters θ_1, θ_2, and α. In our experiments, θ_1 and θ_2 are used to control the shape adversarial loss function. Specifically, θ_1 guides the segmentation network to learn by minimizing the difference between the reconstruction results and the labels, and θ_2 optimizes SAAE by maximizing the difference between the segmentation results and the labels. α is used to control the boundary loss function.

According to Table IV, we found that our SGU-Net can achieve the best performance on the abdominal organ segmentation when $\theta_1 = 5$, $\theta_2 = 0.01$ and $\alpha = 0.5$. It is noted that $\delta = 0.25$ in this experiment. The analysis of the set of δ can be seen in the section of the discussion.

D. Experimental Comparison on Test Datasets

To validate the superiority of the proposed SGU-Net, eleven state-of-the-art networks used for medical image segmentation are considered as comparative approaches. These networks can be roughly grouped into two categories: 2D networks including U-Net [2], R2U-Net [9], Attention U-Net [5], DenseU-Net [16], ResU-Net [15], ResU-Net++ [17], CE-Net [20], U-Net++ [3], and Non-local U-Net [18], as well as 3D networks including 3D U-Net [22] and V-Net [10].

CT liver segmentation: Quantitative metrics on the CHAOS dataset are shown in Table V. We can see that one of the main reasons is that the 3D convolutional architecture is too complex to be used for small datasets. Also, the lower contrast and resolution of the CT images in this dataset, combined with the higher spacing make it difficult to extract the temporal domain information better and may even have the opposite effect.

TABLE V
QUANTITATIVE RESULTS FOR LIVER SEGMENTATION ON THE CHAOS-CT TESTING SET

Method	CHAOS-CT-Liver		
	DICE (%)↑	ASSD (mm)↓	MSSD (mm)↓
U-Net [2]	94.04±2.32	1.70±0.94	29.52±12.23
R2U-Net [9]	94.13±2.20	1.64±0.58	29.50±12.20
Attention U-Net [5]	94.23±2.12	1.55±0.55	29.42±15.0
DenseU-Net [16]	94.15±2.07	1.60±0.90	29.12±14.28
ResU-Net [15]	94.05±2.38	1.71±0.82	29.50±13.12
ResU-Net++ [17]	94.10±2.13	1.65±0.53	29.37±12.35
Non-local U-Net [18]	94.26±2.06	1.54±0.62	29.06±12.27
CE-Net [20]	94.56±1.87	1.50±0.49	28.43±12.13
U-Net++ [3]	93.96±2.03	2.13±0.88	29.88±13.26
3D U-Net [22]	91.62±4.02	2.61±1.20	34.43±22.38
V-Net [10]	92.23±3.29	2.48±1.06	34.21±21.06
SGU-Net	**94.68±1.64**	**1.43±0.35**	**26.26±11.9**

The best values are in bold.

TABLE VI
QUANTITATIVE RESULTS FOR LIVER SEGMENTATION ON THE LITS TESTING SET

Method	LiTS-CT-Liver		
	DICE (%)↑	ASSD (mm)↓	MSSD (mm)↓
U-Net [2]	93.99±1.23	5.79±0.53	90.25±6.28
R2U-Net [9]	94.02±1.56	5.21±0.62	45.92±7.53
Attention U-Net [5]	94.09±1.43	3.42±0.51	42.24±6.35
DenseU-Net [16]	94.11±1.42	4.26±0.56	35.65±6.21
ResU-Net [15]	94.04±1.62	5.74±0.82	61.89±8.28
ResU-Net++ [17]	94.10±1.89	4.36±0.73	32.70±5.87
Non-local U-Net [18]	94.18±1.13	3.64±0.62	31.93±5.96
CE-Net [20]	94.04±1.06	4.11±0.51	51.22±5.82
U-Net++ [3]	94.01±1.18	5.23±0.45	43.26±5.03
3D U-Net [22]	94.10±1.06	2.61±0.43	36.43±5.38
V-Net [10]	94.25±1.03	2.48±0.38	38.28±5.05
SGU-Net	**95.90±1.08**	**1.30±0.21**	**24.45±4.36**

The best values are in bold.

In terms of the segmentation performance of the 2D network, CE-Net tends to be a better solution due to the adoption of pre-trained encoder, and the Dice value improves by 0.52% compared to U-Net. The strategies of residual connectivity, dense connectivity, and recurrent connectivity are also useful, and thus ResU-Net, DenseU-Net, and R2U-Net provide higer segmentation accuracy than U-Net. The setting of attention mechanism brings effective gains, both Attention U-Net as well as ResU-Net++ use spatial attention and channel attention, and the Dice metric is thus improved by 0.19% and 0.06%, respectively, compared to U-Net. Non-local module can obtain global attention by using a larger receptive field, so Non-local U-net shows better performance. We note that SGU-Net obtains the best values of Dice (94.68%), ASSD (1.43 mm), and MSSD (26.26 mm). Similarly, Table VI shows quantitative results for liver segmentation on the LiTS dataset.The LiTS is a larger 3D dataset where a fair comparison with the methods of 3D CNNs

TABLE VII
QUANTITATIVE RESULTS FOR MULTI-ORGAN MR T2 MODITY SEGMENTATION

Method	CHAOS-MRT2-Liver			CHAOS-MRT2-Right Kidney			CHAOS-MRT2-Left Kidney			CHAOS-MRT2-Spleen		
	ASSD (mm)↓	MSSD (mm)↓	DICE (%)↑	ASSD (mm)↓	MSSD (mm)↓	DICE (%)↑	ASSD (mm)↓	MSSD (mm)↓	DICE (%)↑	ASSD (mm)↓	MSSD (mm)↓	DICE (%)↑
U-Net [2]	3.83±2.78	53.56±29.0	90.32±7.41	3.71±4.53	28.83±22.78	88.15±13.1	2.10±1.35	36.18±23.6	90.31±3.46	2.03±1.87	23.45±16.14	89.32±5.82
R2U-Net [9]	3.60±2.80	52.40±31.6	91.75±3.18	1.43±1.14	18.02±10.42	91.68±3.70	1.73±1.40	26.90±23.5	91.52±2.53	2.05±2.33	21.87±17.22	89.84±6.50
Attention U-Net [5]	2.93±2.21	43.86±20.8	92.43±2.68	1.28±1.19	17.16±9.87	92.47±3.98	1.84±2.66	24.30±22.15	92.10±3.01	1.97±1.64	23.21±13.83	89.74±5.13
DenseU-Net [16]	2.74±1.31	35.02±12.76	92.03±2.70	1.52±1.63	16.65±9.25	91.97±4.02	1.60±1.35	22.16±36.5	92.13±3.28	1.17±1.21	16.27±12.65	91.23±5.06
ResU-Net [15]	2.55±1.21	36.40±14.8	91.97±2.62	1.76±1.89	18.01±13.5	91.17±4.42	1.65±**1.02**	23.49±**20.23**	91.82±4.62	2.04±2.32	22.01±15.2	89.33±6.51
ResU-Net++ [17]	2.39±1.44	34.61±14.1	92.40±2.79	1.27±1.14	15.28±8.77	92.12±3.97	1.57±1.51	23.62±22.25	92.60±2.38	1.12±0.95	15.37±11.0	92.29±4.04
Non-local U-Net [18]	1.52±0.95	31.25±16.8	92.68±1.70	1.08±1.15	12.46±7.87	92.78±4.17	1.57±2.08	23.65±22.91	92.67±2.97	1.87±1.52	21.06±11.28	91.27±5.16
CE-Net [20]	1.57±0.99	27.30±**13.2**	93.69±**1.68**	1.01±1.06	15.87±10.1	93.02±3.72	1.68±2.90	23.90±23.52	92.79±3.30	1.07±0.98	11.61±7.84	92.40±3.54
U-Net++ [3]	4.42±3.76	60.58±31.28	90.20±5.23	1.67±1.29	28.85±22.80	89.32±7.01	1.67±1.03	35.14±22.37	89.91±4.64	2.17±2.32	23.16±17.4	88.26±5.57
3D U-Net [22]	14.23±4.01	70.26±33.58	65.23±8.36	10.03±5.26	65.87±50.2	64.17±7.08	9.26±4.76	50.78±33.51	61.28±13.2	15.14±10.05	89.20±67.32	50.66±21.3
V-Net [10]	12.13±3.68	66.87±32.3	67.83±7.62	8.06±4.03	62.32±47.2	67.72±6.96	7.80±3.62	48.04±36.5	64.83±11.2	16.2±9.96	91.46±68.2	49.36±20.3
SGU-Net	**1.48±0.87**	**25.49**±14.12	**94.77**±1.92	**0.91±1.04**	**11.23±7.61**	**93.32±3.45**	**1.32**±1.27	**21.04**±22.63	**93.05±2.13**	**0.80±0.42**	**10.87±3.48**	**93.26±3.26**

The best values are in bold.

baseline can be made. We can see that SGU-Net provides 1.8% and 1.65% improvement in Dice compared to 3D U-Net and V-Net, respectively. Also, SGU-Net provides finer segmentation contours, with the lowest ASSD (1.30 mm) and MSSD (24.45 mm) compared to other networks.

The above experimental results indicate that SGU-Net can make the prediction map achieve better contour and shape consistency, and its ability to mimic expert annotations performs significantly better.

Abdominal multi-organ MR segmentation: Table VII shows quantitative results for multi-organ MR T2 modity segmentation. As for 3D networks, they do not provide the required robustness for organ segmentation. The detailed reason has been presented in the second paragraph of Section IV. D. experimental comparison. For 2D networks, significant improvements can be noticed using attention for right kidney, left kidney and spleen. Attention U-Net (spatial attention), ResU-Net++ (channel attention), and Non-local U-Net (Non-local attention) clearly provide similar effect on liver, left kidney, and right kidney segmentation. On the basis of 89.32% Dice of U-Net, both ResU-Net++ and Non-local U-Net show greater improvement (+2.97%, +1.95%) on spleen segmentation than Attention U-Net (+0.42%). DenseU-Net improves the value of Dice by 0.06%, 0.8%, 0.31% and 1.9% compared to ResU-Net on liver, left kidney, right kidney and spleen, respectively.

It is easy to see that the contribution of CE-Net is clear. Using pre-trained models can provide better underlying features and multi-scale feature extraction blocks can extract richer features, thus CE-Net provides better Dice, ASSD and MSSD values than U-Net. Our proposed SGU-Net obtains the best segmentation results, not only in the Dice metric, but also in the ASSD and MSSD metrics for the four types of organs, which indicates that our strategy is able to constrain the organ contour leading to higher segmentation accuracy.

E. Efficiency Analysis

Table VIII reports a comparison of parameters, FLOPS (floating point operations), model size and Dice of different models on the CHAOS-CT. FLOPS and Memory are estimated with an input size of $1 \times 256 \times 256$. Compared to 2D networks, 3D networks require more memory and higher computational costs due to the use of 3D convolutional kernels. For 2D network, depthwise separable convolution in mobileNet can effectively reduce the model parameters, and the number of model parameters decreases by 82.65% compared to that of U-Net. Furthermore, we can clearly see that SGU-Net is superior to comparative models since it achieves the highest segmentation accuracy but requires only 4.99 M parameters, 19.56 MB model size, and 4.98 GFLOPs of computation.

TABLE VIII
COMPARISON OF THE EFFICIENCIES OF DIFFERENT NETWORKS

Method	Para. (M)↓	GFLOPS↓	ModelSize (MB)↓	Dice %↑
U-Net [2]	34.53	65.47	121.33	94.04
R2U-Net [9]	39.09	153.01	152.78	94.13
Attention U-Net [5]	34.88	66.59	136.34	94.23
DenseU-Net [16]	12.26	8.83	48.70	94.15
ResU-Net [15]	13.04	80.89	51.01	94.05
ResU-Net++ [17]	14.48	70.98	56.67	94.10
CE-Net [20]	29.00	8.91	113.42	94.56
U-Net++ [3]	9.04	33.83	35.34	93.96
Non-local U-Net [18]	7.91	28.57	30.92	94.26
MobileNet [12]	5.99	14.14	23.46	93.17
Ghost-Net [42]	16.94	31.84	66.28	93.84
3D U-Net [22]	40.32	1032.80	143.52	91.62
V-Net [10]	65.17	516.12	248.69	92.23
SGU-Net	**4.99**	**4.98**	**19.56**	**94.68**

The best values are in bold.

V. DISCUSSION

A. Compacting Model Design

The asymmetric convolution usually perform a direct factorization of the $k \times k$ vanilla convolution into $k \times 1$ and $1 \times k$ convolutions. This direct factorization may cause the consequence of significant information loss leading to performance degration of models. There are two possible reasons. First, deep neural networks usually have distributed eigenvalues in them, and they usually rank higher than 1 in practical applications, so the direct decomposition may cause information loss. Secondly, since standard convolution in feature extraction usually uses square convolutional kernels for spatial and channel features co-extraction, asymmetric convolutional kernels may corrupt feature extraction in channel dimension, resulting in chaotic channel feature extraction and thus information loss. Therefore,

the asymmetric convolution in [47] is fused into the square convolution kernel to enrich feature extraction and representation, but the number of parameters remains unchanged.

Unlike the previous work, the proposed ultralight convolution performs feature extraction in space and channel separately by first performing asymmetric convolution channel-by-channel in the spatial dimension, and then performing feature extraction and combination in channels. This avoids to a certain extent the problem of confusing channel feature extraction when asymmetric convolution is directly operated on space and channel. In addition, the channel-by-channel asymmetric convolution allows better feature extraction for abdominal organs with irregular shapes information in space, which is helpful for improving abdominal image segmentation.

B. Extension on 3D Volume Segmentation

To demonstrate the proposed UC is useful for different networks. we discussed the extension of UC to 3D networks. In fact, there are some works [40], [41] that have extended the depthwise separable convolution to 3D CNNs and applied it to medical image segmentation to reduce the number of parameters. The ultralight convolution integrates the advantages of both asymmetric convolution and depthwise separable convolution and thus can be extended to 3D CNNs to further reduce the number of parameters. To evaluate the performance of ultralight convolution on liver segmentation tasks, we consider the 3D Image Reconstruction for Comparison of Algorithm and DataBase (3Dircadb)[1] as experimental data. The dataset is split into 17 patients for training and 5 patients for testing.

As for the experiments on 3D networks, we first pre-trained our network on the LiTS dataset, and then fine-tuned the network on the 3Dircadb dataset. There are two reasons for this. First, different datasets correspond to different collection environments and parameter configurations, and training a network on datasets of different scales can improve the robustness of the network. Second, the LiTS is a large dataset while the 3Dircadb is a small dataset. Using a large dataset for model training, and then using a small dataset for fine-tuning will not only speed up the training efficiency on the small dataset but also avoid the risk of overfitting.

The 3D network consists of two stages namely encoder and decoder. The volume data size used for 3D network input is fixed to $128 \times 128 \times 64$ voxels. The encoder consists of five stages, each of which corresponds to a different image resolution. The 3D Ultralight V-Net network employs ultralight convolution to achieve feature extraction, which is composed of $3 \times 3 \times 1$, $3 \times 1 \times 3$ depthwise asymmetric convolutions and $1 \times 1 \times 1$ pointwise convolution. Following feature extraction, a $2 \times 2 \times 2$ transposed convolution is used for upsampling, and the output image size of the last stage of the encoder is $8 \times 8 \times 4$. The decoder is an inverse process of the encoder. By using the deconvolution operation, the final output feature map size is the same as the input image size.

Table IX reports the comparison of the efficiency of different networks on the 3Dircadb. Compared with 2D CNNs, 3D CNNs obtains a certain increase in segmentation performance, but they require more memory usage and high computational costs. Clearly, 3D ultralight convolution can effectively overcome the shortcomings of 3D CNNs. By using ultralight convolution in V-Net, the number of model parameters, computational costs, and storage usage are drastically reduced. Moreover, the proposed ultralight convolution provides more competitive results than depthwise separable convolution. In addition, it is clear that there is a slight loss in Dice compared to V-Net. Compared to the reducing of training parameters of 98%, the Dice value is only lost by 0.38%. On this basis, we can try to improve segmentation accuracy by exploring some other means liking deep supervision, like what was done in [40], [41], [53], [54]. This is a direction worth exploring.

TABLE IX
COMPARISON OF THE EFFICIENCY OF DIFFERENT NETWORKS ON THE 3DIRCADB DATASET

Method	Para. (M)↓	GFLOPS ↓	ModelSize (MB)↓	Dice %↑
U-Net [2]	34.53	65.47	121.33	92.30
3D U-Net [22]	40.32	1032.8	143.52	94.13
V-Net [10]	65.17	516.12	248.69	**94.56**
MobileNet+V-Net [10] [12]	1.65	58.07	6.56	94.26
3D Ultralight V-Net	**0.86**	**30.27**	**3.46**	94.20

The best values are in bold.

C. Shape-Guided Exploration

For the design of the shape adversarial autoencoder (SAAE), the inspiration comes mainly from [25], [28]. There are four main types of methods used for shape constraint in medical image segmentation.

The first is to use a discriminator [25] to perform binary classification by determining whether the predicted image is a label or not, which mainly constrains the segmentation network by positive and negative signals. Although this approach is useful for improving network performance, the constraint is insufficient and lacks interpretability.

The second is to apply a pre-trained model [24] to feature extraction and compute the difference between the predicted results and the labels simultaneously and use the difference to constrain the segmentation network. This approach mainly depends on the suitability of the pre-trained model for extracting information from medical images.

The third is to use the autoencoder to learn the reconstruction of the prediction results with labels [28], which forces the autoencoder to learn the features in medical images and use the final reconstruction loss to constrain the segmentation network. This approach is more explanatory compared to the first two methods, but the constraints are limited.

The fourth is to use an autoencoder to improve the segmentation results obtained by the segmentation network. Painchaud et al. [13] used an autoencoder constrained by labels to improve segmentation results, so that the final result can be close to the real segmentation result. This postprocessing operation based on the autoencoder actually helps and corrects the original segmentation network, which cannot directly guide the segmentation network to improve segmentation results.

[1]The dataset is available on http://ircad.fr/research/3d-ircadb-01

TABLE X
ABLATION EXPERIMENTS OF UC AND SAAE ON THE CHAOS-CT DATASET USING DIFFERENT BACKBONE NETWORKS

Model	Vanilla	Ultralight conv ($\delta = 0.25$)	SAAE	Para. (M)	GFLOPs	Model size (MB)	Dice (%)
Attention U-Net [5]	✓			34.88	66.59	136.34	94.23
		✓		2.56	6.04	9.98	94.26
		✓	✓	**2.56**	**6.04**	**9.98**	**94.74**
ResU-Net [15]	✓			13.04	80.89	51.01	94.05
		✓		2.91	22.15	11.2	94.14
		✓	✓	**2.91**	**22.15**	**11.2**	**94.81**
U-Net++ [3]	✓			9.04	33.83	35.34	93.96
		✓		1.45	8.33	5.75	94.07
		✓	✓	**1.45**	**8.33**	**5.75**	**94.35**
ResU-Net++ [17]	✓			14.48	70.98	56.67	94.1
		✓		9.12	28.96	34.7	94.15
		✓	✓	**9.12**	**28.96**	**34.7**	**94.76**
Non-local U-Net [18]	✓			7.91	28.57	30.92	94.26
		✓		1.99	10.28	7.66	94.32
		✓	✓	**1.99**	**10.28**	**7.66**	**94.72**
R2U-Net [9]	✓			39.09	153.01	152.78	94.13
		✓		3.69	12.86	14.2	94.22
		✓	✓	**3.69**	**12.86**	**14.2**	**94.60**

The best values are in bold.

The SAAE designed in this paper is based on the third approach, which has been further explored. As the shape of the organ contains high-dimensional information, it is difficult to measure the shape difference between the prediction result and the label directly. To solve the problem, SAAE uses the auto-encoder to explore the representation of the shape in low-dimensional manifold. Since the organs in the labels only contain shape and location information, the SAAE can better encode the shape and location information of the organs after reconstructing a large number of labels and prediction results. As a result, the segmentation network is constrained by using adversarial learning for the difference between prediction results and labels in low-dimensional manifold. Obviously, SAAE is more efficient and more explanatory.

D. UC and SAAE Adaptability

In this section, we performed ablation experiments of UC and SAAE based on different backbone networks to show the universality of our contributions. As shown in Table X, ultralight convolution can reduce the number of model parameters by up to 90% with no reduction in segmentation accuracy. SAAE can further improve the segmentation accuracy without increasing the model parameters. In other words, they greatly reduce the model parameters while providing competitive segmentation results. In conclusion, the two contributions presented in this paper can be applied to different backbone networks as shown in Table X to reduce the number of parameters, balancing the model size and segmentation accuracy.

E. The Analysis of the Thinning Exponent

The hyperparameter δ is the thinning exponent in the proposed SGU-Net. We evaluated the results obtained by SGU-Net using different values of δ. As shown in Table XI, as the value of δ increases, the number of model parameters becomes larger and the segmentation accuracy becomes higher than before. In this paper, SGU-Net with $\delta = 0.25$ can attain the most competitive performance and achieves the best balance between model size and segmentation accuracy.

TABLE XI
ABLATION EXPERIMENTS OF THE HYPERPARAMETER δ IN SGU-NET ON THE CHAOS-CT DATASET

Model	δ	Para. (M)	GFLOPs	Model size (MB)	Dice(%)
SGU-Net	0.25	4.99	4.98	19.56	94.68±1.64
	0.5	7.19	9.65	28.18	94.72±1.72
	0.75	9.42	14.42	36.88	94.74±1.68
	1	11.67	19.31	45.65	94.75±1.60

In Table XI, we can see that although the larger value of δ will lead to higher segmentation accuracy, the computational efficiency of the network will be lower. Therefore, in practical applications, we need to adjust the value of δ according to different task requirements. If the network is deployed on low resource devices, then we should choose a smaller value of δ to improve computational efficiency. In contrast, if we do not consider the problem of computational costs, we can choose a larger value of δ to achieve higher segmentation accuracy.

VI. CONCLUSION

In this paper, we have proposed a shape-guided ultralight network for medical image segmentation. First, an ultralight convolution is presented to factorize vanilla convolution into deepwise asymmetric convolution and pointwise convolution, which integrates the advantages of both asymmetric convolution and depthwise separable convolution. Secondly, a shape-guided module is presented to use the priori knowledge of fixed organ position and shape to constrain the segmentation network to produce results that are closer to the true organ shape. Extensive experiments on the LiTS and the CHAOS have shown that the proposed SGU-Net provides a general and effective solution to achieve high-quality segmentation results in the case of limited memory and computation resources.

REFERENCES

[1] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Image.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[3] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[4] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, "Modified U-Net (mU-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images," *IEEE Trans. Med. Image.*, vol. 39, no. 5, pp. 1316–1325, May 2019.

[5] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," Apr. 2018, *arXiv:1804.03999*.

[6] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," Feb. 2021, *arXiv:2102.04306*.

[7] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," May 2021, *arXiv:2105.05537*.

[8] A. E. Kavur et al., "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101950.

[9] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (r2U-Net) for medical image segmentation," Feb. 2018, *arXiv:1802.06955*.

[10] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE 4th Int. Conf. 3D Vis*, 2016, pp. 565–571.

[11] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proc. ACM Multimedia Asia*, 2019, pp. 1–6.

[12] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," Apr. 2017, *arXiv:1704.04861*.

[13] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin, "Cardiac segmentation with strong anatomical guarantees," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3703–3713, Nov. 2020.

[14] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, and E. P. Xing, "Scan: Structure correcting adversarial network for organ segmentation in chest X-Rays," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer, 2018, pp. 263–273.

[15] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-UNet for high-quality retina vessel segmentation," in *Proc. IEEE 9th Int. Conf. Inf. Technol. Med. Educ.*, 2018, pp. 327–331.

[16] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 568–576, Feb. 2020.

[17] D. Jha et al., "ResUNet : An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia*, 2019, pp. 225–2255.

[18] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local U-Nets for biomedical image segmentation.," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6315–6322.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[20] Z. Gu et al., "Ce-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[21] T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu, and A. K. Nandi, "DefED-Net: Deformable encoder-decoder network for liver and liver tumor segmentation," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 6, no. 1, pp. 68–78, Jan. 2021, doi: 10.1109/TRPMS.2021.3059780.

[22] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2016, pp. 424–432.

[23] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Image.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[24] A. Mosinska, P. Marquez-Neila, M. Koziński, and P. Fua, "Beyond the pixel-wise loss for topology-aware delineation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3136–3145.

[25] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2020, pp. 552–561.

[26] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, and A. K. Nandi, "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *IEEE Trans. Med. Imag.*, early access, Nov. 30, 2022, doi: 10.1109/TMI.2022.3225687.

[27] S. M. R. Al Arif, K. Knapp, and G. Slabaugh, "SpNet: Shape prediction using a fully convolutional neural network," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2018, pp. 430–439.

[28] F. Liu, Y. Xia, D. Yang, A. L. Yuille, and D. Xu, "An alarm system for segmentation algorithm based on shape model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10652–10661.

[29] F. Liu, L. Xie, Y. Xia, E. Fishman, and A. Yuille, "Joint shape representation and classification for detecting PDAC," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2019, pp. 212–220.

[30] R. Yu et al., "NISP: Pruning networks using neuron importance score propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9194–9203.

[31] J.-H. Luo, J. Wu, and W. Lin, "ThiNet: A filter level pruning method for deep neural network compression," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5058–5066.

[32] X. Ding et al., "ResRep: Lossless CNN pruning via decoupling remembering and forgetting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4510–4520.

[33] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-Net: ImageNet classification using binary convolutional neural networks," in *Eur. Conf. Comput. Vis.*, 2016, pp. 525–542.

[34] Y. Liu, W. Zhang, and J. Wang, "Zero-shot adversarial quantization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1512–1521.

[35] H. Chen et al., "Data-free learning of student networks," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3514–3522.

[36] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Adv. Neural Inf. Process. Syst.*, 2018.

[37] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3048–3068, Jun. 2022.

[38] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[40] T. Lei, W. Zhou, Y. Zhang, R. Wang, H. Meng, and A. K. Nandi, "Lightweight V-Net for liver segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 1379–1383.

[41] J. Zhang, Y. Xie, P. Zhang, H. Chen, Y. Xia, and C. Shen, "Light-weight hybrid convolutional network for liver tumor segmentation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4271–4277.

[42] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1580–1589.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[44] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[45] H. Park, Y. Yoo, G. Seo, D. Han, S. Yun, and N. Kwak, "Concentrated-comprehensive convolutions for lightweight semantic segmentation," 2018, *arXiv:1812.04920*.

[46] Y. Li, X. Li, C. Xiao, H. Li, and W. Zhang, "EACNet: Enhanced asymmetric convolution for real-time semantic segmentation," *IEEE Signal Process. Lett.*, vol. 28, pp. 234–238, 2021.

[47] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1911–1920.

[48] W. Shi et al., "Is the deconvolution layer the same as a convolutional layer?," Sep. 2016, *arXiv:1609.07009*.

[49] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng, "Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks," *Med. Phys.*, vol. 45, no. 10, pp. 4558–4567, 2018.

[50] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2019, pp. 285–296.

[51] P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," *Med. Image Anal.*, vol. 84, Feb. 2023, Art. no. 102680.

[52] H. R. Roth et al., "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 556–564.

[53] Y. Liu, M.-M. Cheng, D.-P. Fan, L. Zhang, J.-W. Bian, and D. Tao, "Semantic edge detection with diverse deep supervision," *Int. J. Comput. Vis.*, vol. 130, no. 1, pp. 179–198, 2022.

[54] S. Mishra, Y. Zhang, D. Z. Chen, and X. S. Hu, "Data-driven deep supervision for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1560–1574, Jun. 2022.