MobileViTv2-ResUNet: U-shaped transformer architecture for precise retinal vessel segmentation

Yu Liu, Yuqing Fu*, Tingxi Wen*
College of Engineering, Huaqiao University, Quanzhou 362021, Fujian, China

ABSTRACT

Over the years, U-Net has become a predominant model in the domain of retinal vessel image segmentation. However, its constrained receptive field and the inherent biases associated with convolutional operations present significant challenges in effectively capturing long-range dependencies. In recent years, although Transformer-based techniques have been integrated into the U-Net architecture to overcome this limitation, the self-attention mechanism inherent in Transformers demands substantial computational resources, thereby increasing computational complexity and the risk of overfitting. To address these challenges, we propose a model that integrates lightweight Transformer and CNN networks, namely MobileViTv2-ResUNet, for precise retinal vessel segmentation. We chose U-Net as the framework for the automated retinal vessel segmentation model. Firstly, in the encoding phase, we introduced MobileViTv2 blocks to replace traditional convolutional modules for feature extraction. Subsequently, inverted residuals are employed within the encoding phase to perform downsampling operations, thereby reducing computational complexity while enhancing the network's representation and generalization capabilities. Additionally, an ASPP module is incorporated between the encoder and decoder to effectively fuse feature information from different scales. Finally, in the decoding phase, we integrate our designed LeakyRes module to prevent the occurrence of the "neuron death" phenomenon, thereby improving the accuracy of retinal vessel segmentation. We validated our MobileViTv2-ResUNet on the public datasets HRF and STARE. Experimental results demonstrate that our MobileViTv2-ResUNet outperforms most existing state-of-the-art algorithms, significantly enhancing vessel segmentation methods, particularly for images with anomalies, bifurcations, and microvessel segmentation challenges.

Keywords: Retinal vessel image segmentation, MobileViTv2 block, Inverted Residuals, LeakyRes

1. INTRODUCTION

Retinal vascular images contain rich geometric structures, such as vessel diameters, branching angles, and lengths, which ophthalmologists can use to prevent and diagnose diseases such as hypertension, diabetes, and atherosclerosis[1]. However, the intricate topology of retinal vessels makes manual segmentation not only labor-intensive and time-consuming but also susceptible to subjective factors. Therefore, automatic retinal vessel segmentation technology has become a research hotspot in the field of medical imaging to meet the practical diagnostic needs.

Early medical image segmentation methods were primarily based on contours and traditional machine learning algorithms[2,3]. With the development of deep Convolutional Neural Networks (CNNs), U-Net[4] was introduced for medical image segmentation. Due to its simple U-shaped structure and superior performance, various U-Net-like methods have emerged, such as Res-UNet[5], Dense-U-Net[6], U-Net++[7], U-Net3+[8], 3D U-Net[9], and V-Net[10]. These CNN-based methods have shown excellent performance in retinal vessel segmentation, demonstrating the strong feature learning capabilities of CNNs.

Although CNN-based methods have achieved outstanding results in medical image segmentation, they still cannot fully meet the strict requirements for segmentation accuracy in medical applications. Due to the inherent locality of convolutional operations, CNN-based methods often struggle to learn explicit global and long-range semantic information interaction[11]. In recent years, researchers have discovered that Transformers can address the inductive bias of locality in CNN networks, enabling them to establish non-local relationships more effectively. The combination of both can provide abundant local information representation for global modeling. Consequently, efforts are underway to integrate CNNs with Transformers to address the shortcomings of CNNs in medical image segmentation. Chen et al.

*fuyq@hqu.edu.cn

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133951I · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048337

pioneered TransU-Net[11], a model that combines Transformers with CNNs, demonstrating the effectiveness of Transformers as medical image segmentation encoders. Subsequently, numerous researchers have leveraged the complementarity of Transformers and CNNs to enhance model segmentation capabilities, such as Medical Transformer[12], TransFuse[13], TransBTS[14], U-NETR[15], and CoTr[16].

Currently, most models combining CNNs and Transformers use Transformer modules centered on self-attention mechanisms. This results in a significant increase in computational complexity and a high risk of overfitting. Additionally, these models have relatively simple structures, making them insufficient for handling complex vascular distributions and achieving precise vascular segmentation. Moreover, retinal datasets are more scarce compared to other medical datasets, failing to meet the extensive training data requirements of Transformer models.

To address these issues, we propose a retinal vessel segmentation model that combines a lightweight Transformer with a CNN (MobileViTv2-ResUNet). Firstly, considering that U-Net can achieve good segmentation performance even with limited medical datasets, we chose U-Net as the CNN framework for the retinal vessel segmentation model. We then replaced the convolutional modules of the U-Net encoder with lightweight Transformer modules (MobileViTv2 blocks) for feature extraction. Simultaneously, we incorporated Inverted Residuals into the encoder for downsampling. This not only deepens the network structure but also prevents the "vanishing gradient" problem, thereby enhancing the model's ability to handle complex vascular distributions and achieve precise vascular segmentation. Next, we integrated Atrous Spatial Pyramid Pooling (ASPP) between the encoder and decoder as a bridge, expanding the effective receptive field to encompass a broader context. In the decoder part, we designed the LeakyRes module for feature reconstruction, preventing the "neuron death" phenomenon and further improving the precision of retinal vessel segmentation. Additionally, we added auxiliary heads in the decoder, providing extra loss functions and introducing gradients earlier in the network, thereby aiding the primary task learning and further enhancing model performance. Extensive experiments demonstrate that this method exhibits good segmentation accuracy and robust generalization capabilities on retinal vessel datasets.

2. METHOD

2.1 Architecture overview

The overall architecture of our proposed MobileViTv2-ResUNet is depicted in Figure 1. MobileViTv2-ResUNet comprises an encoder, ASPP (Atrous Spatial Pyramid Pooling), decoder, skip connections, and auxiliary head. Below, we will delve into the details of each block.

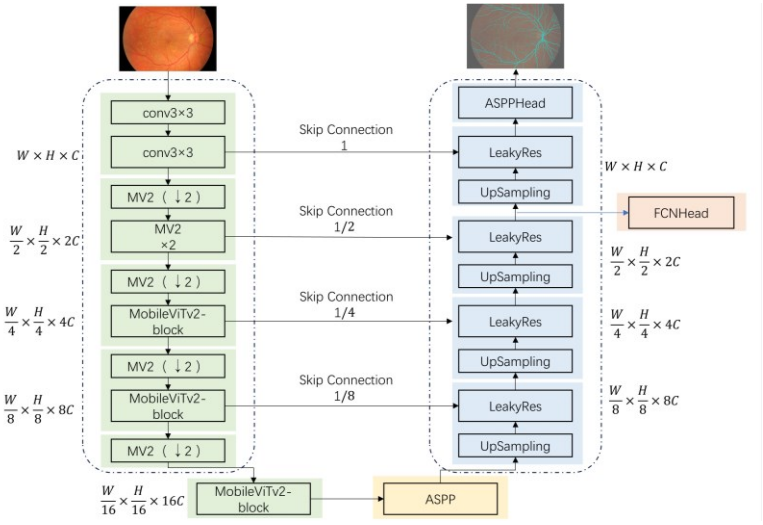

Figure 1. Architecture of MobileViTv2-ResUNet.

2.2 Encoder

In the encoder, we first input the retinal fundus image with a resolution of (H, W, 3) into a 3×3 convolutional layer, projecting the feature dimension to an arbitrary dimension (denoted as C), and generate a feature map with a resolution of (H, W, C). Next, we apply an Inverted Residual Block (stride=2) (denoted as MV2(↓2)) to achieve 2x downsampling

and increase the feature dimension to twice the original dimension, resulting in a feature map of (H/2, W/2, 2C). Then, we use two Inverted Residual Blocks (stride=1) (denoted as MV2) to increase the network depth, thereby enhancing the feature representation capability, which helps in solving complex tasks. Finally, we apply a combination of Inverted Residual Blocks (stride=2) and MobileViTv2 Blocks three times, effectively extracting both local and global features while downsampling, enhancing the feature representation capability, and improving the model's performance in retinal image segmentation.

Inverted Residual (stride=2) structure: This structure executes downsampling to reduce the resolution of the feature map while increasing the receptive field to better capture global features and contextual information in the image.

Inverted Residual (stride=1) structure: This structure maintains the size of the feature map while still introducing nonlinear transformations and feature extraction, thereby enhancing the model's expressive power.

2.3 ASPP

In the proposed architecture, we not only use ASPP as a bridge between the encoder and decoder but also employ it as the decoding head of this model, as shown in Figure 1. We will detail the ASPP structure serving as the bridge in this section, while the ASPP decoding head will be discussed in Section 3.4.

The ASPP module in this section receives input feature maps of size (H/16, W/16, 16C), and through convolution blocks with dilation rates of 1, 6, 12, and 18, respectively, it captures features at different scales. Then, these features are element-wise summed to obtain a feature map fused with multiscale feature information. Finally, the fused features undergo additional convolutional layers to maintain the output feature map size as (H/16, W/16, 16C).

2.4 Decoder

The decoder consists of four repeated decoding layers and the ASPP decoding head. Each decoding layer first performs interpolation convolution upsampling (InterpConv) to restore the image's resolution, then the fused feature map, passed through the skip connection, is inputted into our designed LeakyRes module for information propagation. The design of the LeakyRes module aims to prevent gradient vanishing, thus facilitating better information propagation, as illustrated in Figure 2. We chose LeakyReLU[17] as the activation function. The functionality of LeakyReLU is very similar to ReLU, with the only difference being that when the input is less than zero, the output of the ReLU activation function is zero, whereas in LeakyReLU, the part of the input less than zero is not zero but is multiplied by a small slope (usually referred to as the leak parameter) to allow small negative gradients to pass through. The mathematical expression of LeakyReLU is shown in equation (1).

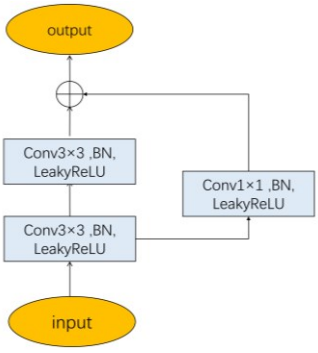

Figure 2. LeakyRes module.

$$f(x)=\begin{cases} x, x \geq 0 \\ \alpha x, x < 0 \end{cases} \tag{1}$$

where α is a small slope, typically taken as a very small positive number, such as 0.01.

Compared to ReLU, the main advantage of LeakyReLU is that it can avoid the problem of dead neurons, where some neurons have gradients that are always zero during training, causing those neurons to have zero response to the input. Therefore, we use LeakyReLU as the activation function for the intermediate layers to circumvent this issue.

2.5 Skip connection

We use skip connections to merge the multi-scale features from the encoder part with the upsampled features from the decoder part, and feed the fused features into the LeakyRes module. Skip connections help the network better capture features from different levels, thereby improving segmentation accuracy and preserving details.

2.6 Auxiliary head

In the third decoder layer, an FCN auxiliary head is added, which takes the feature map output from the third layer of the decoder as input and converts it into a segmentation mask used only for computing the auxiliary loss. By this point, with the fusion and upsampling through multiple decoder layers, the input feature map already contains rich semantic information. The FCN auxiliary head generates additional segmentation predictions, which are compared with the output of the ASPP decoding head and used to calculate additional CrossEntropyLoss loss functions. These additional loss functions provide extra supervision signals, helping to improve the model's performance and robustness.

3. EXPERIMENTS

To evaluate the MobileViTv2-ResUNet architecture, we used two publicly available retinal vessel segmentation datasets to train, validate, and test the model. We compared the performance of our MobileViTv2-ResUNet model with the current state-of-the-art methods.

3.1 Dataset and preprocessing

The HRF dataset consists of 45 high-resolution fundus images with a resolution of 2336×3504 pixels, divided into 15 subsets. Each subset contains one healthy fundus image, one image of a patient with diabetic retinopathy, and one image of glaucoma. Since no segmentation for training and testing sets is provided, we used the first 5 subsets as the training set and the remaining 10 subsets for evaluation. The STARE dataset comprises 20 fundus images with a resolution of 700x605 pixels, including 10 healthy fundus images and 10 images with pathological features. We manually divided the STARE dataset into training and testing images in a 10/10 ratio.

We performed image preprocessing using PhotoMetric Distortion to simulate various lighting conditions and camera settings that may occur during fundus image acquisition, thereby improving the model's adaptability and robustness to real-world situations. Considering the limited annotated training samples on HRF and STARE and the absence of available pretrained weights, we adopted rotation and flipping data augmentation methods to prevent overfitting and enhance segmentation accuracy.

3.2 Evaluation metrics

This paper uses five metrics, including Intersection over Union (IoU), Accuracy, F1-score, Precision, and Recall, to evaluate the effectiveness of the algorithm in segmenting retinal images.

3.3 Implementation details

MobileViTv2-ResUNet is implemented based on Python 3.8 and PyTorch 1.13.1. For the HRF dataset, we set the crop_size of input images to 256×256 and the stride to 170. For the STARE dataset, we set the crop_size of input images to 128×128 and the stride to 85. We train our model on an NVIDIA GeForce RTX 3070 Ti Laptop GPU with 16GB of memory. We use the SGD optimizer with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005 to optimize our model. In terms of learning strategy, we adopt a polynomial learning rate scheduler, which gradually decreases the learning rate from the initial value to the minimum value of 1×10^{-4} in a polynomial manner. During training, we use an iteration-based training loop, with a maximum of 20000 iterations, and perform validation every 2000 iterations to monitor the model's performance during training.

3.4 Experiment results on HRF dataset

We trained the proposed MobileViTv2-ResUNet model on the HRF dataset and compared it with the current state-of-the-art methods. The comparison results of the metrics are shown in Table 1.

Table 1. Comparison of different methods on HRF.

Model	IoU	Accuracy	F1-score	Recall	Precision
U-Net	65.26	73.9	78.98	73.9	84.81
MobileNetV2+FCN	64.08	73.47	78.11	73.47	83.37
Res-UNet	65.23	75.1	78.22	75.1	84.92
CCNet	64.62	72.99	78.51	72.99	**84.93**
ANN	65.64	74.41	79.26	74.41	84.79
Ours	**66.99**	**77.2**	**80.23**	**77.2**	84.05

The experimental results indicate that our MobileViTv2-ResUNet consistently achieves the highest IoU, F1-score, Accuracy, and Recall across all benchmark tests. The compared methods include U-Net[4], MobileNetV2[18]+FCN (using MobileNetV2 network as backbone with FCNHead as decoding head), Res-UNet[5], CCNet[19], and ANN[20]. The segmentation results of different methods on the HRF dataset are illustrated in Figure 3. From Figure 3, it can be observed that CNN-based methods are prone to over-segmentation issues, which may be attributed to the locality of convolutional operations. In contrast, our proposed model obtains segmentation results most similar to the Ground truth, demonstrating the superior generalization ability and robustness of our approach.

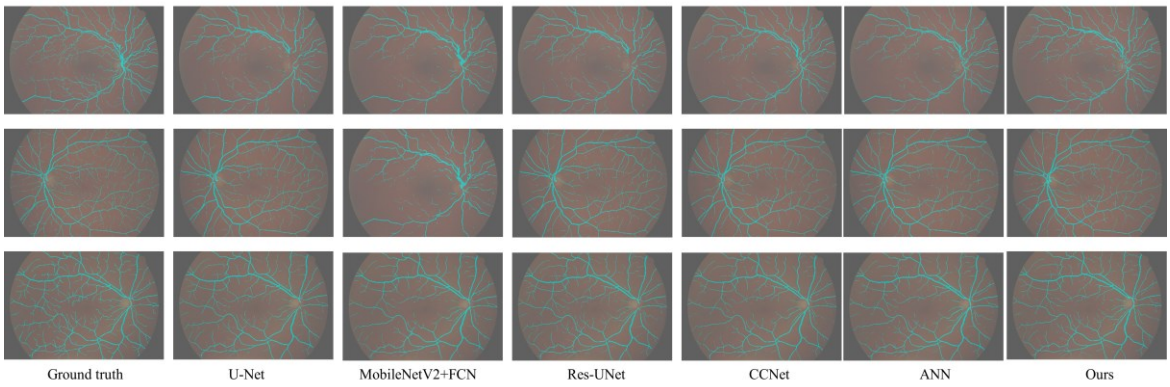

Figure 3. Segmentation results of different methods on HRF dataset.

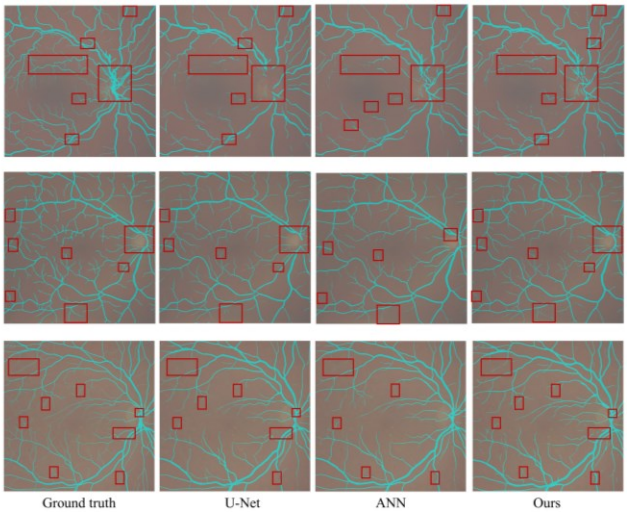

Figure 4. Segmentation detail comparison of U-Net, ANN, and MobileViTv2-ResUNet.

Figure 4 provides a visual comparison of U-Net, ANN, and MobileViTv2-ResUNet on HRF images. From Figure 4, we observe that the segmentation results of U-Net and ANN exhibit fewer clear edges of retinal vessels, with some noise and discontinuities present. These noises may arise from the instability of the models when dealing with small vessels. In contrast, the segmentation results of MobileViTv2-ResUNet demonstrate clearer and finer edges of retinal vessels, indicating its capability to better capture the shape and structure of vessels.

1.1 3.5 Experiment results on STARE dataset

The training results of MobileViTv2-ResUNet on the STARE dataset are presented in Table 2. We observed that although CCNet and ANN performed well on the high-resolution HRF dataset, their segmentation results significantly degraded on the low-resolution STARE dataset. This is attributed to the rich details in high-resolution images, enabling context-aware networks to effectively learn complex structures, whereas the lack of details and reduced pixels in low-resolution images allow U-Net-based networks to more accurately recover and maintain local image features. Despite the attention mechanism introduced by CCNet and ANN, they are still less effective than the U-shaped architecture in handling subtle local features and small-scale details. Additionally, MobileViTv2-ResUNet continues to perform excellently on low-resolution images, achieving an accuracy of 77.2%, further demonstrating the superior generalization ability and robustness of our approach.

Table 2. Comparison of different methods on STARE.

Model	IoU	Accuracy	F1-score	Recall	Precision
U-Net	68.14	77.82	81.5	77.82	84.57
MobileNetV2+FCN	45.82	59.93	62.84	59.93	66.06
Res-U-Net	69.02	78.19	81.67	78.19	**85.48**
CCNet	46.48	60.01	63.46	60.01	67.34
ANN	46.62	59.95	63.59	59.95	67.70
Ours	**69.57**	**79.67**	**82.06**	**79.67**	84.59

The segmentation results of different methods on the STARE dataset are shown in Figure 5. From Figure 5, it can be observed that although U-shaped CNN-based architectures are less effective in handling details compared to our MobileViTv2-ResUNet when processing low-resolution images, they outperform non-U-shaped architectures like MobileNetV2+FCN in terms of segmentation performance. This further confirms the correctness of our choice to use U-Net as the framework for our retinal vessel segmentation model.

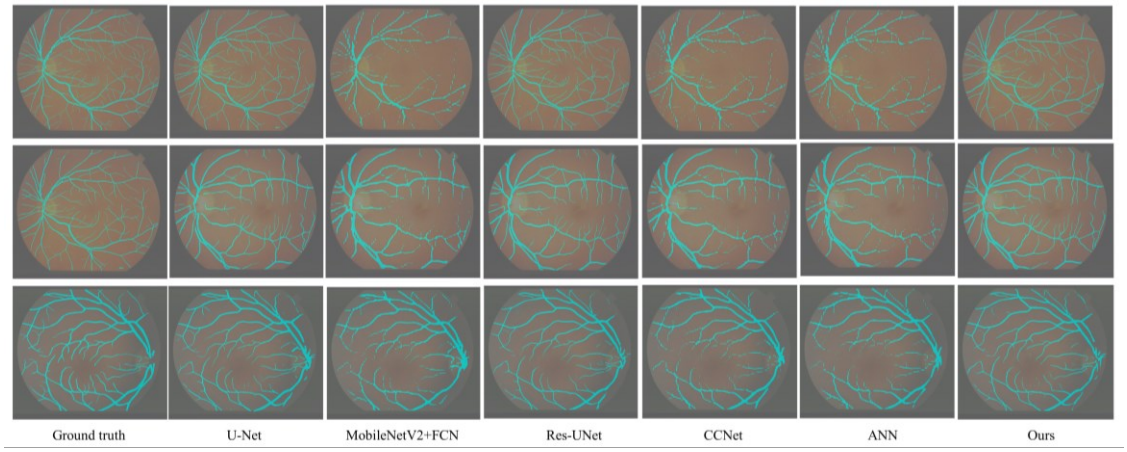

Figure 5. Segmentation results of different methods on STARE dataset.

4. CONCLUSION AND OUTLOOK

This paper proposes a novel Transformer-based U-shaped encoder-decoder for retinal vessel segmentation. We incorporate the MobileViTv2 block to the U-shaped encoder-decoder structure, leveraging the powerful capabilities of Transformers to model global dependencies, thereby better capturing long-range semantic correlations in the retinal vessel segmentation task. To enhance the network's performance in segmenting complex and fine vessels, we utilize Inverted Residuals and the LeakyRes module designed within this model, effectively addressing the "vanishing gradient" problem when deepening the network. Inverted Residuals (with a stride of 2) also replace the downsampling layers in U-Net, reducing model parameters. Unlike U-Net, we introduce auxiliary heads in the decoder, which provide additional loss functions and introduce backpropagated gradients earlier in the network, thereby assisting the primary task learning and further enhancing model performance.

We validate the effectiveness of MobileViTv2-ResUNet on two public datasets, HRF and STARE. Compared to some state-of-the-art methods, MobileViTv2-ResUNet not only achieves the highest F1-score, IoU, Accuracy, and Recall, but also performs excellently in segmenting fine vascular structures. MobileViTv2-ResUNet provides more accurate and detailed information on retinal vessel distribution, which is expected to assist ophthalmologists in diagnosing and planning treatment for eye diseases.

Currently, the segmentation accuracy of MobileViTv2-ResUNet in densely vascularized areas requires further improvement. In future research, we plan to optimize the network structure further. Additionally, we will explore the use of medical image datasets from different modalities for cross-modal learning, such as combining fundus images with optical coherence tomography (OCT) images, to enhance the model's adaptability and generalization capabilities.

REFERENCES

[1] Chen, Y., Jiang, Y., Yuan, Y., et al., "Retinal microvascular segmentation algorithm based on multi-scale attention mechanism," 2022 International Conference on Virtual Reality, Human-Computer Interaction and Artificial Intelligence (VRHCIAI), 72-77 (2022).

[2] Tsai, A., Yezzi, A., Wells, W., et al., "A shape-based approach to the segmentation of medical imagery using level sets," IEEE Transactions on Medical Imaging 22(2), 137-154 (2003).

[3] Held, K., Kops, E. R., Krause, B. J., et al., "Markov random field segmentation of brain MR images," IEEE Transactions on Medical Imaging 16(6), 878-886 (1997).

[4] Ronneberger, O., Fischer, P. and Brox, T., "U-Net: Convolutional networks for biomedical image segmentation," Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015, 234-241 (2015).

[5] Xiao, X., Lian, S., Luo, Z., et al., "Weighted Res-UNet for high-quality retina vessel segmentation," 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 327-331 (2018).

[6] Wu, Y., Wu, J., Jin, S., et al., "Dense-U-net: Dense encoder–decoder network for holographic imaging of 3D particle fields,"Optics Communications 493, 126970 (2021).

[7] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., et al., "UNet++: A nested U-Net architecture for medical image segmentation," Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 3-11 (2018).

[8] Huang, H., Lin, L., Tong, R., et al., "UNet 3+: A full-scale connected unet for medical image segmentation," ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1055-1059 (2020).

[9] Çiçek, Ö., Abdulkadir, A., et al., 3"D U-Net: Learning dense volumetric segmentation from sparse annotation," Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016, 424-432 (2016).

[10] Milletari, F., Navab, N. and Ahmadi, S. A., "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," arXiv 1606.04797, (2016).

[11] Chen, J., Lu, Y., Yu, Q., et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv 2102.04306, (2021).

[12] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., et al., "Medical transformer: Gated axial-attention for medical image segmentation," arXiv 2102.10662, (2021).

[13] Zhang, Y., Liu, H. and Hu, Q., "TransFuse: Fusing transformers and CNNs for medical image segmentation," arXiv 2102.08005, (2021).

[14] Wang, W., Chen, C., Ding, M., et al., “TransBTS: Multimodal brain tumor segmentation using transformer,” arXiv 2103.04430, (2021).
[15] Hatamizadeh, A., Tang, Y., Nath, V., et al., “UNETR: Transformers for 3D medical image segmentation,” arXiv 2103.10504, (2021).
[16] Jiang, W., Trulls, E., Hosang, J., et al., “COTR: Correspondence transformer for matching across images,” arXiv 2103.14167, (2021).
[17] Liu, Y., Wang, X., Wang, L., et al., “A modified leaky ReLU scheme (MLRS) for topology optimization with multiple materials,”Applied Mathematics and Computation 352, 188-204 (2019).
[18] Sandler, M., Howard, A., Zhu, M., et al., “MobileNetV2: Inverted residuals and linear bottlenecks,” arXiv 1801.04381, (2018).
[19] Huang, Z., Wang, X., Wei, Y., et al., “CCNet: Criss-cross attention for semantic segmentation,” arXiv 1811.11721, (2018).
[20] Zhu, Z., Xu, M., Bai, S., et al., “Asymmetric non-local neural networks for semantic segmentation,” arXiv 1908.07678, (2019).