IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 June 2023, accepted 24 July 2023, date of publication 27 July 2023, date of current version 2 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299491

RESEARCH ARTICLE

BSANet: High-Performance 3D Medical Image Segmentation

QI HUANG[1], JUN SU[1], KRZYSZTOF PRZYSTUPA[2], AND OREST KOCHAN[1,3]
[1]School of Computer Science, Hubei University of Technology, Wuhan 430068, China
[2]Department of Automation, Lublin University of Technology, 20-618 Lublin, Poland
[3]Department of Information-Measuring Technologies, Lviv Polytechnic National University, 79013 Lviv, Ukraine

Corresponding author: Qi Huang (huangqi@hbut.edu.cn)

This work was supported by the Artificial Intelligence Laboratory of the Hubei University of Technology.

ABSTRACT As a challenge in the field of smart medicine, medical picture segmentation gives important decisions and is the basis for future diagnosis by doctors. In the past decade, FCN-based network topologies have made amazing progress in the field. However, the limited perceptual capacity of convolutional kernels in FCN network topologies limits the network's ability to acquire a global field of view. We propose BSANet, a 3D medical image segmentation network based on self-focus and multi-scale information fusion with a high-performance feature extraction module. BSANet can help the network to extract deeper features by obtaining a larger range of perceptual capabilities by using its self-focus and multi-scale information aggregation pooling modules. Brain tumor segmentation dataset and multi-organ segmentation dataset are used to train and evaluate our model. BSANet produces excellent results with its high-performance feature extraction network with an attention module and multi-scale information fusion module.

INDEX TERMS Deep learning, FCN, medical image segmentation.

I. INTRODUCTION

In the context of intelligent medicine, medical picture segmentation can be considered as a semantic segmentation technique application. One of the most difficult jobs in the world of computer vision is the segmentation of lesions, backdrops, and human organ tissues from a medical image. This task necessitates the separation of each pixel. Due to the ongoing advancement of deep learning techniques, methodologies based on neural networks and deep learning have recently gained popularity and acceptance as a mainstream way of studying medical image segmentation. For the purpose of semantic segmentation, Long et al. [1] proposed the FCN network in 2014 and got excellent results. This network's encoder-decoder structure proposal has had a significant influence on the creation of the following versions. In order to combine low- and high-resolution feature maps and effectively fuse low- and high-resolution image features, Ronneberger et al. [2] proposed the U-Net network model for medical image segmentation. U-Net has since established itself as the industry standard for the majority of medical image segmentation tasks. Since the majority of medical data, including CT and MRI images, are 3D data in practice, Çiçek et al. [3] proposed the 3DU-Net model using a 3D convolution kernel to better mine the high latitude spatial correlation of the data. Currently, 3DU-Net has become the mainstream basic architecture in the field of medical image separation. Compared to 3D-UNet, V-Net [4] uses residual connectivity design [5] for a deeper network (4 sub-sampling) to higher performance.

The MultiResUnet network was proposed by Ibtehaz and Rahman [6] In order to extract spatial features at various scales while reducing the computational effort required by the network, MultiResUnet modifies the convolutional blocks and jump connections in U-Net using the concept of residuals. U-Net uses a series of 3×3 convolutional kernels to simulate the perceptual field of 5×5 convolutional kernels and 7×7 convolutional kernels.

Cascade models are typically trained with two or more models for image segmentation tasks to increase segmentation accuracy. The segmentation of medical images using this technique is particularly common.

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva.

Three categories—coarse-fine segmentation, detection, segmentation, and hybrid segmentation—can be used to classify cascade models. The first type is the coarse-fine segmentation framework, which employs a cascade of two 2D networks for segmentation. Using the findings of the first network's coarse segmentation, the second network model implements fine segmentation. For the segmentation of the liver and hepatic tumors, Christ et al. [7] suggested a cascade network. The network initially utilized one FCN to segment the liver, and then it fed the results of the prior liver segmentation into a second FCN to segment the liver tumor. In order to quickly and coarsely segment the liver of the entire image of a CT volume, Tang et al. [8] trained a straightforward convolutional-deconvolutional neural network (CDNN) model (19-layer FCN). They then applied a second CDNN (29-layer FCN) to the liver region to segment the liver in a more precise manner. This cascaded network efficiently extracts richer multi-scale contextual information than a typical cascaded network by using the posterior probability produced by the initial network.

While 2D convolutional neural networks are unable to learn the temporal information in 3D and 3D convolutional neural networks are frequently computationally expensive and have significant GPU memory consumption, medical images are primarily composed of 3D body data. Consequently, a few pseudo-3D segmentation techniques are suggested. In order to forecast the core slice, Vu et al. [9] employed the superposition of nearby slices as input. The generated 2D feature maps were then fed into a conventional 2D network for model training. Due to the use of only local temporal information, these pseudo-3D algorithms can segment targets from 3D data but have limited accuracy improvement. 2D and 3D cascade networks perform better than pseudo-3D networks. For the segmentation of the liver and hepatic tumors, Li et al. [10] developed a hybrid densely linked U-Net (H-DenseUNet). Using a simple ResNet, the technique first achieves rough results for liver segmentation [5], extracts 2D image features effectively with a 2D DenseUNet, then extracts 3D image features effectively with a 3D DenseUNet, and finally constructs a hybrid feature fusion layer for joint optimization of 2D and 3D features. The model is still complex, and the 3D convolution still has a lot of parameters, even if H-DenseUNet reduces the complexity of the model relative to the entire 3D network. Similar in structure to the H-DenseUNet, the lightweight hybrid convolutional network (LW-HCN) suggested by Zhang et al. [11] uses 3D depth-separable convolution, which requires fewer parameters and lower processing costs.

It is typically difficult to recognize tiny anatomical features with fuzzy noise borders using traditional UNet [2]. to deal with this issue. To conduct brain interlayer segmentation, Valanarasu et al. [12]suggested the overcomplete cascade network KiU-Net. By adding an upsampling layer after each conversion layer of the encoder, the authors were able to create a novel overcomplete structure called Ki-Net in which the middle layer's spatial size is greater than the spatial size of the input data. In order to increase the overall segmentation accuracy, the suggested Ki-Net cascades with U-Net and has a stronger edge capture capability than U-Net. Along with improving segmentation accuracy, Ki-Net's low-level fine edge feature map and U-Net's high-level shape feature map enable quick convergence for small anatomical markers and fuzzily noisy boundaries.

In order to build a particular kind of convolutional neural network, dense connections are frequently used. Each layer's input in densely connected networks is taken from the outputs of all preceding layers. Instead of each U-Net sub-block, Guan et al. [13] proposed a modified U-Net made up of dense connections. Although dense connection aids in obtaining richer picture features, it also tends to increase the number of parameters and somewhat lower the robustness of the feature representation. All of the U-Net layers (from one to four layers) were connected by Zhou et al. [14] This topology has the benefit of enabling the network to automatically learn the significance of features at various levels. Additionally, the jump connections were changed to enable the decoder to aggregate features with various semantic scales, creating a highly adaptable feature fusion approach. The utilization of thick connections continues to have the drawback of increasing the number of parameters. In order to decrease the number of parameters, a pruning procedure is incorporated into the model optimization.

Deep networks typically outperform shallow networks for CNNs, but they also have certain new issues such as gradient disappearance, challenging network convergence, and high memory footprints. Inception solves these issues. Better performance is obtained by merging convolutional kernels in parallel rather than deepening the network. Using multi-scale convolutional kernels, the structure may extract more complex visual features and combine those elements to provide a better feature representation. Gu et al. [15] 's CE-Net proposal involved integrating Inception into the segmentation of medical images. However, Inception is typically complicated, which makes it challenging to modify the model.

Encoder-decoder-based structures have the drawback that when downsampling, a certain amount of information is lost that cannot be made up by upsampling. Hu et al. [16] suggested a channel-based attention module to address this issue. This module calculates the weights of each channel in parallel in the convolution operation branch, suppressing the channels with trivial features, and emphasizing the channels with important characteristics.

A brand-new attention mechanism built on local space and combining channel attention was put forth by Woo [17]. In order to segment medical images, they compared the effectiveness of channel attention, spatial attention, and various combinations of the three types of attention. They then proposed the CBAM (solution block attention module), a lightweight attention mechanism with both local space and

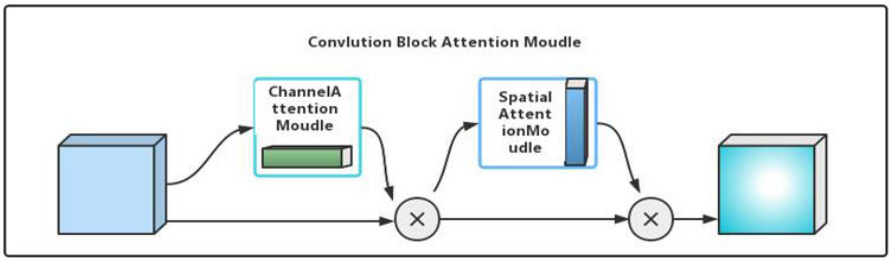


FIGURE 1. Structure of CBAM attention module.

channel suppression that can be quickly applied to a variety of convolutional neural network models.

In 2022, Wang [18] proposed yolo_v7, which, thanks to its accurate and lightweight encoder, outperforms yolo_v4 in terms of accuracy while reducing the network size by approximately 75%. We modified the YOLO_V7 model to make it work with our 3D feature extraction network, increasing network accuracy and lowering network parameter count.

Since full-volume neural networks (FCNs) [1] were proposed for the division field, U-type symmetrical structural networking has become mainstream in the field of medical imaging, and the proposal of U-Net [2] has inspired a lot of creative transformation and exploration, including the modification of network coders such as V-Net [4], 3D-UNet [3], etc., using jump-connection guidelines for sampling information recovery, such as U-NET++ [13], AttentionU-Net [19], RA-UNEet [20] and so on. Currently, the mainstream approach is to combine the attention mechanisms such as Transformers [21], SE-Block [15], CBAM [17] and U-net [2], such as TransU-net [22], Swin-Unet [23], and other methods to introduce the improvement in network performance, which is obvious, but also brings an increase in the number of network parameters, so we chose the lightweight attention module of CBAM to be embedded in our division network. Figure 1 illustrates the structure of the CBAM attention module.

More and more models are trained utilizing multimodal fusion due to the peculiarity of medical data formats, such as Dense Multi-path U-Net [24], Cascaded U-Net [25], etc., which employ multimodal training and so on to mine the deep information between various modalities. Medical image segmentation networks' performance can also be enhanced by adding multi-scale feature fusion modules.Polar Transformation M-Net [26], Focal Tversky Attention U-Net [27], etc.

II. RELATED WORK

The FCN (Full Convolutional Neural Network) architecture is the fundamental CNN architecture for the majority of semantic segmentation problems. It can be thought of as consisting of two encoder and decoder components, where the encoder is made up of a stack of convolutional and pooling layers and the decoder is made up of an inverse convolutional and convolutional layer. When compared to the basic CNN architecture, the FCN [1] architecture omits the linear layer and is more effective because it doesn't have to deal with issues with duplicate storage and computational convolution because it uses pixel blocks instead.

It adds a jump layer to the FCN so that the encoder's data can be utilized to direct the decoder's information recovery, which boosts the network's speed. U-Net [2] is a network topology created expressly for medical picture segmentation. In order to further boost the network's performance, Res-U-Net [5] uses a jump layer in the convolutional layer inside the encoder and decoder. On the other hand, multiResUNet [6] is expressly created with a ResPath rather than a jump connection.

Based on the DeepLab architecture, Chen et al. introduced the DeepLabV3+ architecture [28]. The DeepLab technique [29] employs a null convolutional sum (ASPP) (Atrous Spatial Pyramid Pooling) to tackle the segmentation problem and a conditional random field (CRF) model for post-processing. DeepLab, unlike FCN and U-Net, does not focus on the symmetry of the structure, and this two-branch synthesis structure can more effectively reduce the computational effort of the network. The encoder of DeepLabV3+ is primarily based on null convolution and ASPP, while the decoder uses simple low-level and deep-level feature synthesis. Figure 2 shows the structure of several networks, and their similarities and differences can be seen.

Where the channel attention module calculation process can be described as the following equation:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))), \end{aligned} \tag{1}$$

the process of calculating the spatial attention module can be described as the following equation:

$$\begin{aligned} M_S(F) &= \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times7}([F^S_{avg}; F^S_{max}])\right). \end{aligned} \tag{2}$$

Figure 2 demonstrates the difference and connection of the four network structures, U-Net network adds hop connection to FCN to enhance information fusion, and Res-U-Net specializes in U-Net by designing ResBlock for downsampling and upsampling.DeepLab V3+ designs ASPP module for

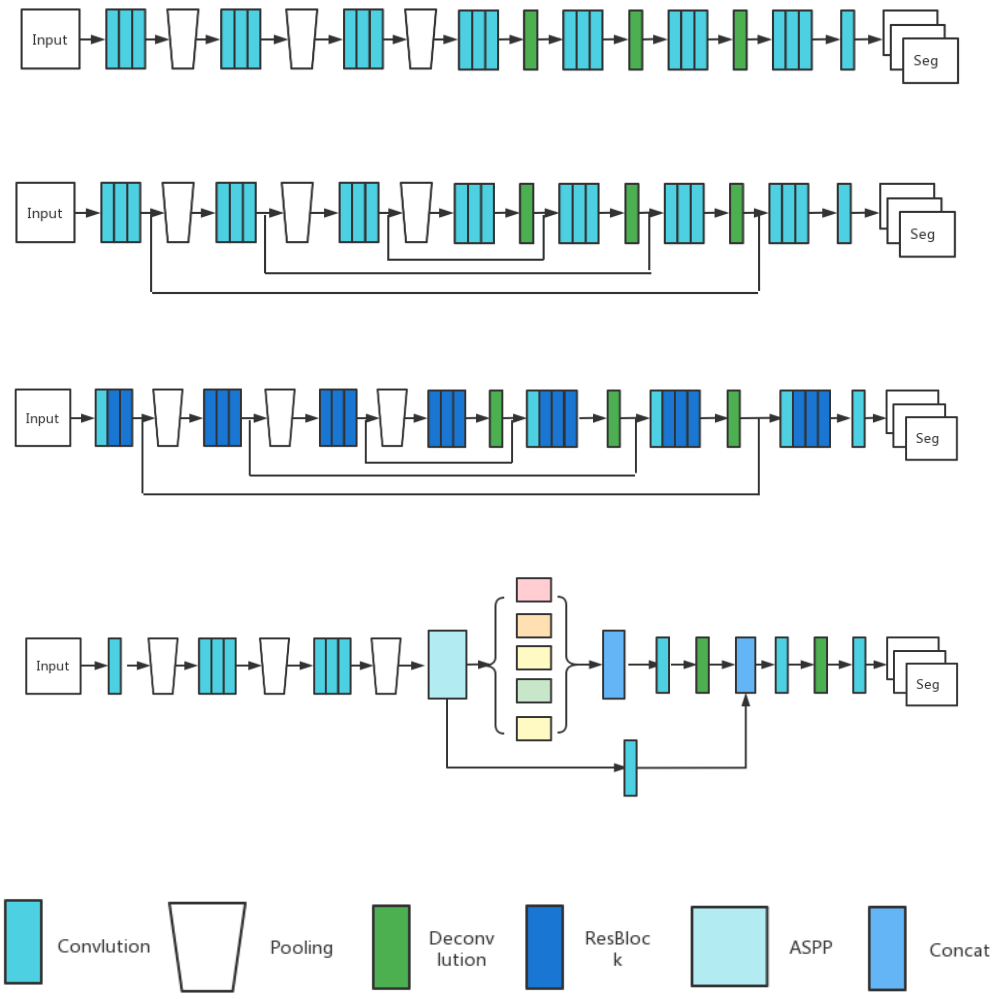


FIGURE 2. Top-down diagram of FCN,U-Net,Res-U-Net,DeepLabV3+ network structure.

image segmentation without pursuing a completely symmetric structure.

III. BSANET

A. OVERVIEW

Figure 3 illustrates the network structure of BSANET, which, from an overall view, consists of two large modules, the encoder module and the decoder module. We use the FCN structure as the most basic framework, incorporating the most advanced encoder and decoder modules in the current computer vision field, and adding a multi-scale feature fusion and attention module, which allows end-to-end training of medical images. We borrowed the high-performance feature extraction module from YOLO_V7 [17] to design a feature extraction network suitable for medical impact segmentation, firstly, the image is passed through the first CBS layer to boost the channel of the image to 32, and then two branches are formed by two downsampling, and after two downsamplings, branch one passes through the 3D-DAPPM [30] multi-scale feature fusion module It reaches the decoder module. Branch II passes through the SPPCSPC pyramid pooling module after two more downsamplings, and then reaches the decoder module after two upsampling. The branch that goes through 3D-DAPPM has low-level features, and the branch that goes through SPCSPC has deep-level features, and after 2 upsampling it is spliced with the low-level feature map of branch one, and then it goes through 3×3 convolution to achieve the effect of feature fusion, and finally it goes through upsampling to get the output.

Each downsampling is followed by a CBAM hybrid attention module, which uses the global attention mechanism and the channel attention mechanism to reduce the loss caused by downsampling.

B. BACKBONE BLOCK

An efficient and accurate feature extraction module is the key to the whole network. By controlling the shortest and longest gradient path, the deeper network can learn and converge

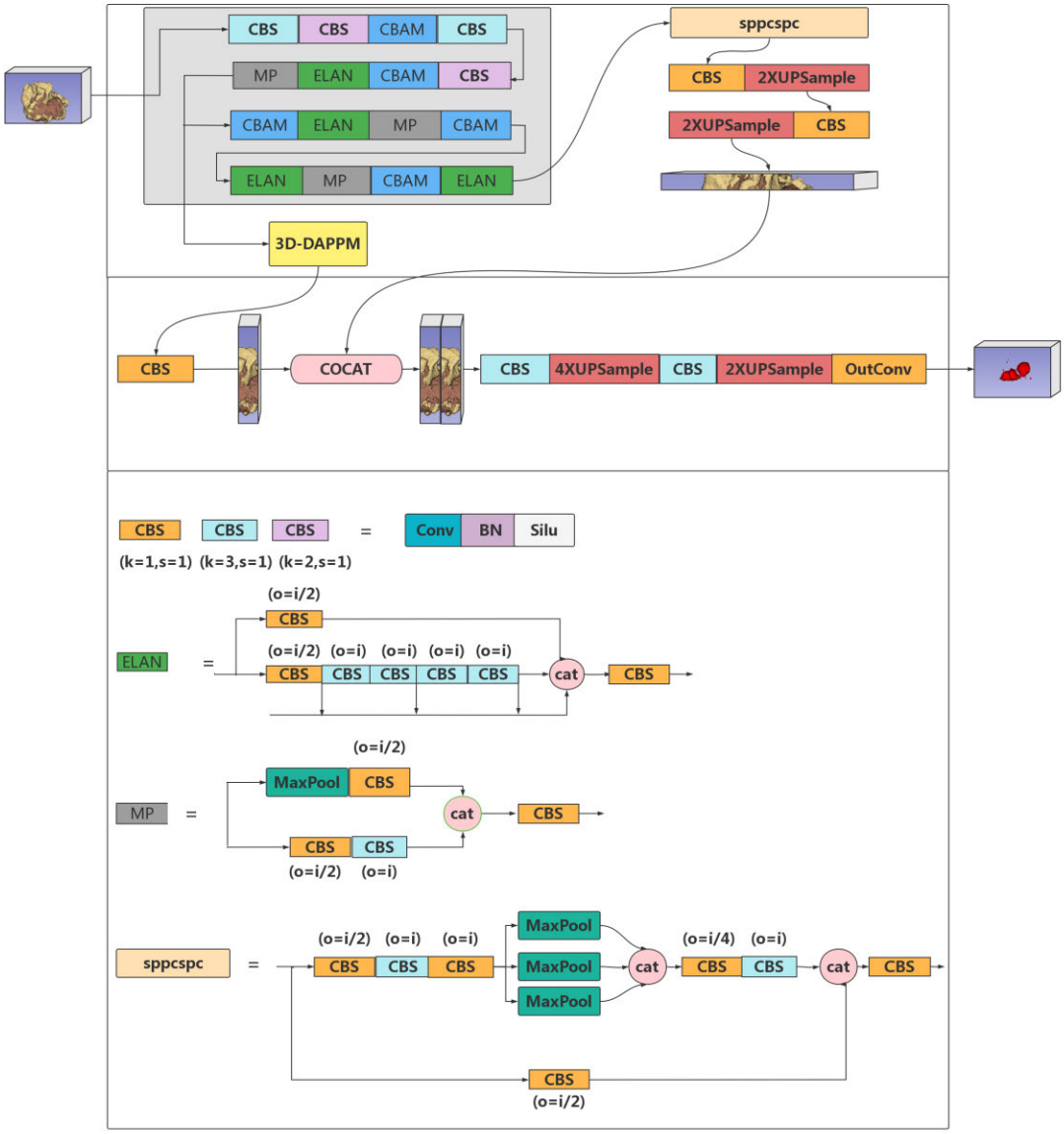


FIGURE 3. BSANet network structure diagram.

efficiently. As shown in Figure 3, ELAN is a four-branch aggregation structure, which learns deeper features by deepening the network while stitching and fusing features from different levels to achieve efficient and accurate feature extraction. Depending on the output dimension of the last convolutional block in the ELAN module, the ELAN module can optionally change the Channel dimension of the changing feature map. The MP module combines the two methods to extract features of different depths, giving full play to the advantages of the two downsampling methods. An ELAN module and an MP module can be regarded as one downsampling module. Unlike FCN and U-Net, which simply combine convolutional layers for downsampling, ELAN and MP modules combine for downsampling without increasing the number of network parameters, the network level is deeper, and shallow features are retained for fusion, which is more efficient and accurate than other networks. sppcspc module is a spatial pyramid pooling aggregation module, and its role is to fuse multi-scale information. It is worth stating that we are inspired by the efficient feature extraction network of YOLO_V7.

C. 3D-DAPPM

We enhanced the DAPPM [30] (Deep Aggregate Pyramid Pooling) module for our network, which is a five-branch structure that uses convolution kernels of various sizes to downsample the input feature maps for 2D semantic segmentation. For 3D convolution, we use a triple convolution. In 3D convolution, we use trilinear interpolation for upsampling, stitch the results from each branch separately to get

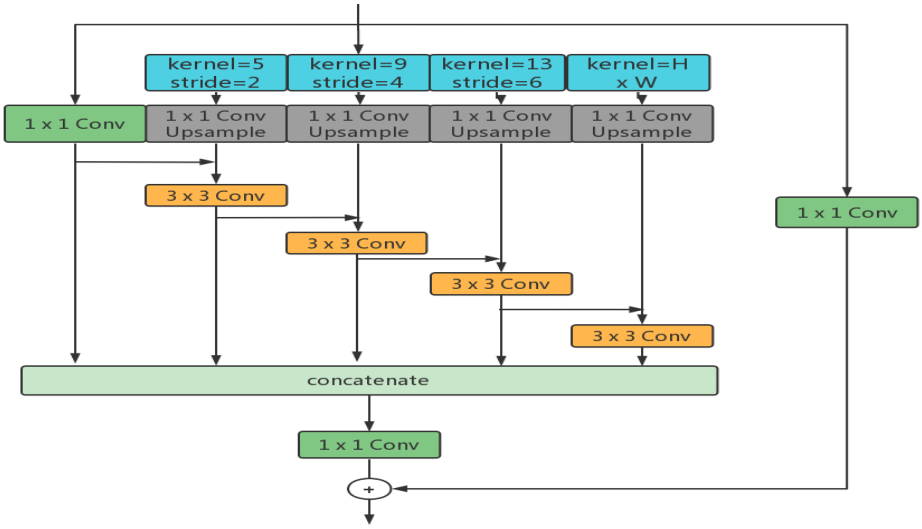


FIGURE 4. Deep aggregation pyramidal pooling detail diagram.

the outputs Y1–Y5, stitch the results from the five branches together in the depth direction, and connect the residuals with the input feature map to achieve feature fusion.

We use the multiscale information to guide the upsampling information recovery by using a quarter size of the input feature map to input to the 3D-DAPPM module and fusing the multiscale data of the shallow sub-feature map with another branch of the higher-level feature map.

As shown in Figure 4, using the input feature maps and picture level data produced by global average pooling, we use a quarter feature map input and employ a large pooling kernel with linear steps to construct 1/8, 1/16, and 1/32 image resolution feature maps, respectively. The y on each branch can be expressed as the following equation for each input x:.

$$y_i = \begin{cases} C_{1\times1}(x), & i = 1 \\ C_{3\times3}\left(U\left(C_{1\times1}\left(p_{4i-3,2i}(x)\right)\right) + y_{i-1}\right), & 1 < i < n \\ C_{3\times3}\left(U\left(C_{1\times1}\left(P_{global}(x)\right)\right) + y_{i-1}\right), & i = n. \end{cases} \tag{3}$$

D. LOSS FUNCTION AND EVALUATION INDEX

The total loss is the sum of dice loss and cross-entropy loss:

$$L_{total} = L_{dice} + L_{CE}, \tag{4}$$

dice loss:

$$L_{dice} = -\frac{2}{|k|}\sum_{k\in K}\frac{\sum_{i\in I} u_i^k v_i^k}{\sum_{i\in I} u_i^k + \sum_{i\in I} v_i^k}, \tag{5}$$

where u is the network's softmax output and v is the split label's one-hot encoding.

An evaluation metric used in segmentation is called Miou. Similar to semantic segmentation, medical picture segmentation needs the classification of each pixel and the right labels and expected results can be thought of as sets. It is possible to determine the accuracy of the algorithm for various segmentation objects by taking the intersection of the two sets, or the number of predicted pairs of pixels, for each category and dividing it by the concurrent set of the two sets, or the number of predicted pairs of pixels plus the number of predicted wrong pixels.

$$MIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{6}$$

IV. EXPERIMENTS AND RESULTS

A. BRAIN TUMOR SEGMENTATION

1) OVERVIEW

Brain tumors are usually diagnosed using multimodal MRI, including four four-modal data: native T1-weighted (T1), contrast T1-weighted (T1-Gd), T2-weighted (T2) and T2 liquid-attenuated inversion recovery (FLAIR) image sequences, and we need to apply algorithms to label and merge three nested subregions in the images: whole tumor (WT), tumor core (TC) and enhancing tumor (ET) segmentation. Brain tumor segmentation has proven to be a very difficult subject.

Experiment A and Experiment B were carried out on a Linux server equipped with an RTX A5000 (24 GB) graphics card and an Intel(R) Xeon(R) Gold 6330 CPU running at 2.00 GHz. It was required to make sure that the GPU graphics RAM was 12 GB or more because a 3D network has more parameters than a 2D network.

2) DATASET

The BRATS2021 dataset, the oldest of all MICCAI contests, served as the basis for our model's training and validation.

TABLE 1. Comparison with different networks of Dice and HD95 evaluation metrics.

Model	DICE				HD95			
	ET	TC	WT	Mean	ET	TC	WT	Mean
U-Net[2]	0.839	0.877	0.907	0.874	11.122	10.243	9.205	10.19
Attention-UNet[19]	0.850	0.877	0.915	0.881	10.447	10.463	9.004	9.971
TransUNet[22]	0.883	0.910	0.926	0.906	10.421	14.501	14.027	12.983
BSANet	**0.886**	**0.926**	**0.936**	**0.916**	**10.221**	**9.845**	**8.876**	**9.647**

TABLE 2. Ablation experimental data.

Model	DICE				MIOU			
	ET	TC	WT	Mean	ET	TC	WT	Mean
Without CBAM	0.877	0.907	0.927	0.903	0.781	0.830	0.864	0.823
Without DAPPM	0.880	0.905	0.926	0.904	0.786	0.826	0.862	0.825
Without C and D	0.855	0.882	0.907	0.881	0.747	0.789	0.830	0.787
BSANet	**0.886**	**0.926**	**0.936**	**0.916**	**0.795**	**0.862**	**0.880**	**0.845**

A comparison of the experimental results without the CBAM attention module and the DAPPM deep aggregation pyramid pooling module in the network structure is shown in respectively.

It is one of the most advanced resources for learning medical picture segmentation and has been used for ten years straight through 2021. The BRATS dataset, which contains a training set (1251 cases), a validation set (219 instances), and a test set (530 cases) of mpMRI scans from a total of 2000 patients, is multi-institutional, multi-parametric, and multimodal. The validation and test sets lack segmentation labels while the training set has both photos and labels for each section.

3) PRE-PROCESSING AND AUGMENTATION

We used picture enhancement methods on the images such as center cropping, random flip, Gaussian noise, contrast and brightness modification. Most of the medical impact datasets are MRI multimodal datasets, so multimodal training provides a better test of the connection between several modalities compared to unimodal training. We wrote images of the same case in 3 modalities with GT images into the same H5 file to synthesize 4D images for training. The segmentation challenge for brain tumors has three categories, four input channels, and the data format is H×W×D×C, where C=4.

4) RESULTS

Figure 5 shows the loss curves of BSANet and U-Net during the training process. We visualized some of the results, as shown in Figure 6. The results are provided in Table 1 using U-Net with Attention-UNet as the baseline criterion. When we examine the convergence of BSANET and Unet during the training process, we can see that the network converges much more quickly than UNet and has a far lower initial loss. We compared the outcomes with those of other widely used models after validating our model on the validation set. Our model performs better than other network models in the test set, according to the experimental results. We also examine the impact of the DAPPM and CBAM attention modules on the network structure. The experiment's findings are displayed in Table 2. The results of the experiments demonstrate that the accuracy of the network structure is positively influenced by both the CBAM attention module and the DAPPM module. Because the attention mechanism will lessen information loss during downsampling and multiscale pyramidal pooling will fuse multiscale information to guide the upsampling recovery, using FCN with attention mechanism and multiscale information fusion model is preferable to using FCN network alone.

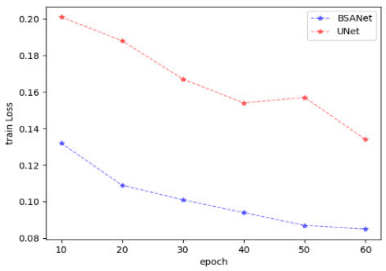


FIGURE 5. Comparison of the loss curves of BSANet and UNet during training.

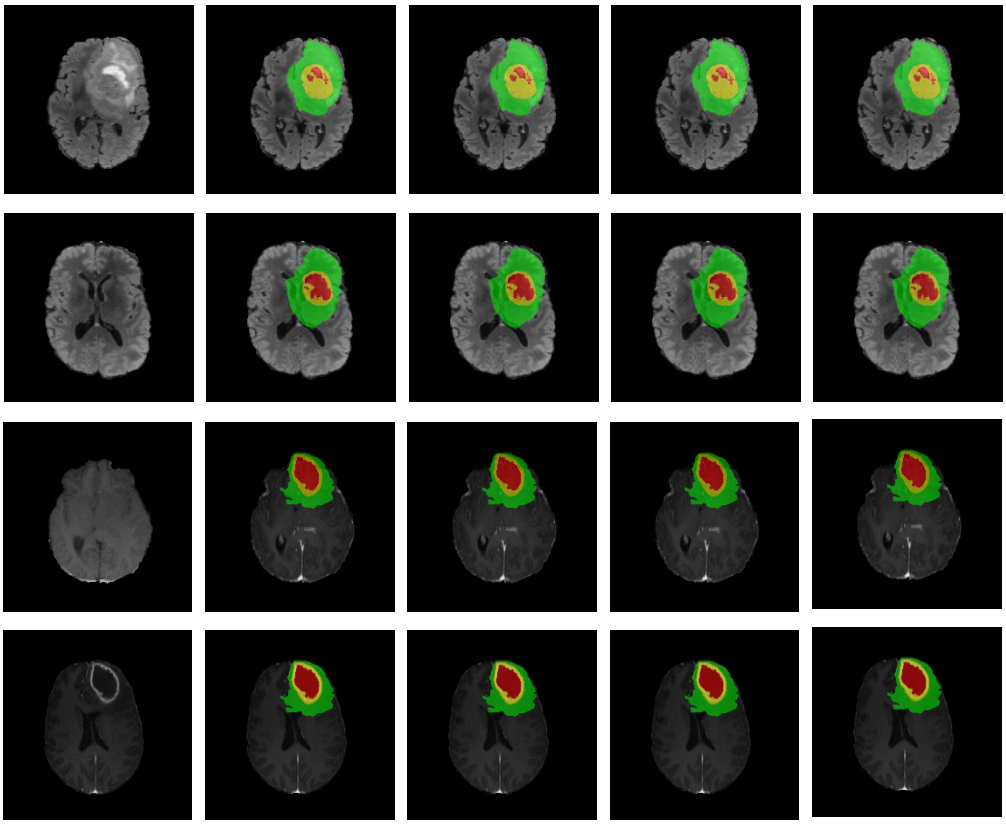


FIGURE 6. Tumor segmentation visualization images.

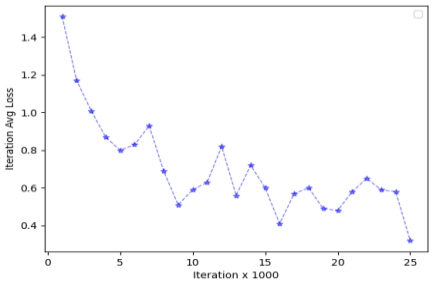


FIGURE 7. The loss curve during the training process.

The first column is the original image, the second column is the GT segmentation map, and the third and fourth columns are the segmentation maps of 3DU-Net, TransUNet and BSANet, respectively.

B. MULTI-ORGAN SEGMENTATION

1) OVERVIEW

Segmenting the abdominal organs is essential for clinical diagnosis since the abdominal organs have some mobile and variable features that cause them to have a diversity of forms and sizes. The difficulty of segmenting the abdominal organs is particularly challenging due to the indistinct demarcation between the abdominal organs. The most common multi-organ segmentation methods today use deep learning techniques.

2) DATASET

The 30 people in the BTCV dataset [31] had their abdomens CT-scanned, and clinical radiologists at Vanderbilt University Medical Center directed interpreters to label 13 organs on those scans. Each volume underwent different preprocessing by having the intensity normalized between [-1000,1000] HU and [0,1]. During the preprocessing phase, every image was resampled to an isotropic voxel spacing of 1.0 mm. The chosen data format was $H \times W \times D \times C$,, where C=1. As a 13-class, 1-channel input segmentation job, multi-organ segmentation was completed and supplied into BSANet.

3) RESULTS

Figure 7 shows the loss curve during training, and Figure 8 shows some of the visualization results. The outcomes are reported in Table 3 using U-Net as our baseline criterion. We still use the FCN network structure and add the lightweight attention mechanism module to the network structure, which reduces the overall network parameters. We also use the 3D network, which is superior to the 2D network in mining the different modalities in medical images, and the experimental results show that our network

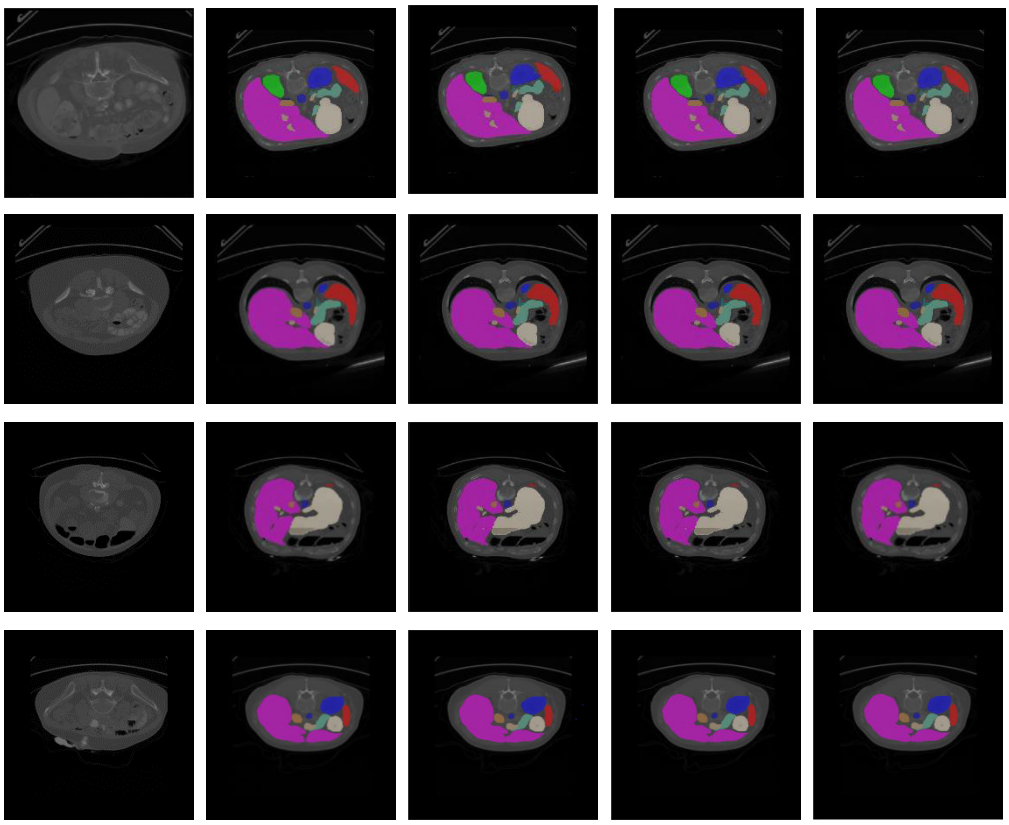


FIGURE 8. Shows the segmentation results of different algorithms on the multi-organ dataset.

TABLE 3. Comparison of the accuracy of different networks on multi-organ segmentation.

Methon	Spl	PKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	AVG
SETR NUP[32]	0.931	0.890	0.897	0.652	0.760	0.952	0.809	0.867	0.745	0.717	0.719	0.620	0.796
SETR PUP[32]	0.929	0.893	0.892	0.649	0.764	0.954	0.822	0.869	0.742	0.715	0.714	0.618	0.797
SETR MLA[32]	0.930	0.889	0.894	0.650	0.762	0.953	0.819	0.872	0.739	0.720	0.716	0.614	0.796
nnUNet[33]	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
ASPP[28]	0.935	0.892	0.914	0.689	0.760	0.953	0.812	0.918	0.807	0.695	0.720	0.629	0.811
TransUNet[22]	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
UNTER[34]	**0.972**	0.942	0.954	0.825	0.864	**0.983**	0.945	0.948	0.890	0.858	**0.799**	0.812	0.891
BSANet	0.967	**0.945**	**0.957**	**0.836**	**0.876**	0.976	**0.956**	**0.966**	**0.902**	**0.866**	0.786	**0.823**	**0.905**

structure outperforms other networks in terms of accuracy on the BTCV dataset. 3D networks have higher accuracy rates than 2D networks because they are better able to mine the spatial information of several imaging modalities in medical images. The network using the Transformer global attention mechanism has better performance than the traditional FCN architecture, but the fusion of information at different scales and depths in the high-performance feature extraction module of BSANet allows the whole network to show superior performance in segmenting organs of different sizes.

The first column shows the original image, the second column shows the GT labels, and the third, fourth, and fifth columns show the segmentation results of U-Net, nn-UNet [32], and BSANet, respectively.

V. DISCUSSION AND CONCLUSION

In this study, we present a deep learning network model with a number of critical features, including a high-performance feature extraction module, an effective decoder module, a multi-scale information fusion module, and an attention module. These modules greatly increase the network's segmentation

accuracy for medical images, are highly flexible and adaptable, and are simple to include in different CNN architectures.

The performance of the CNN network can be improved by designing an attention module with high accuracy and a multi-scale information fusion module. We have conducted experiments on the brain tumor segmentation dataset and the multi-organ segmentation dataset, respectively. On the brain tumor segmentation dataset, we have also conducted comparison experiments. From the comparison experiments, we can see the improvement of each module on the network performance.

Since the majority of medical pictures are 3D structures like CT and MRI, we suggest a network structure based on 3-dimensional segmentation. Medical image segmentation is crucial for clinical purposes. BSANet beats existing 3D medical image segmentation networks now in use in terms of network performance. Deep learning is playing an increasingly visible role in the field of intelligent medicine, and the use of artificial intelligence for medical research is of extraordinary significance. Hopefully, in the near future, our network will play a role in the medical environment, helping doctors to diagnose patients' conditions more accurately and quickly.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[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. Intervent.* Cham, Switzerland: Springer, Nov. 2015, pp. 234–241.

[3] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. 19th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Athens, Greece: Springer, Oct. 2016, pp. 424–432.

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

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[6] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.

[7] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempflfler, M. Armbruster, F. Hofmann, M. D'Anastasi, W. H. Sommer, S.-A. Ahmadi, and B. H. Menze, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2016, pp. 415–423.

[8] W. Tang, D. Zou, S. Yang, and J. Shi, "DSL: Automatic liver segmentation with faster R-CNN and deeplab," in *Proc. Int. Conf. Artif. Neural Netw.*, Sep. 2018, pp. 137–147.

[9] M. H. Vu, G. Grimbergen, T. Nyholm, and T. Löfstedt, "Evaluation of multi-slice inputs to convolutional neural networks for medical image segmentation," 2019, *arXiv:1912.09287*.

[10] 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. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[11] 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. (IJCAI)*, 2019, pp. 4271–4277.

[12] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "KiU-Net: Towards accurate segmentation of biomedical images using over-complete representations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2020, pp. 363–373.

[13] 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 Informat.*, vol. 24, no. 2, pp. 568–576, Feb. 2020.

[14] 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.

[15] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[17] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[18] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[19] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[20] Q. Jin, Z. Meng, C. Sun, H. Cui, and R. Su, "RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans," *Front. Bioeng. Biotechnol.*, vol. 8, Dec. 2020, Art. no. 605132.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[22] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[23] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Computer Vision—ECCV Workshops*. Cham: Cham, Switzerland: Springer, 2023, pp. 205–218.

[24] J. Dolz, I. B. Ayed, and C. Desrosiers, "Dense multi-path U-Net for ischemic stroke lesion segmentation in multiple image modalities," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Granada, Spain: Springer, Jan. 2019, pp. 271–282.

[25] H. Liu, X. Shen, F. Shang, F. Ge, and F. Wang, "CU-Net: Cascaded U-Net with loss weighted sampling for brain tumor segmentation," in *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy*. Shenzhen, China: Springer, Oct. 2019, pp. 102–111.

[26] 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. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[27] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 683–687.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[29] 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. 2018.

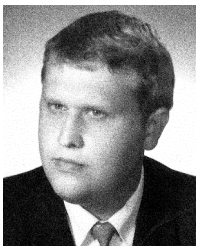
[30] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," 2021, *arXiv:2101.06085*.

[31] T. Tong, R. Wolz, Z. Wang, Q. Gao, K. Misawa, M. Fujiwara, K. Mori, J. V. Hajnal, and D. Rueckert, "Discriminative dictionary learning for abdominal multi-organ segmentation," *Med. Image Anal.*, vol. 23, no. 1, pp. 92–104, Jul. 2015.

[32] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2021, *arXiv:2012.15840*.
[33] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.
[34] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 1748–1758.

QI HUANG is currently pursuing the master's degree in computer science with the Hubei University of Technology. He is a Software Developer and an Open Source Community Contributor working on the application of artificial intelligence in the medical field.

JUN SU received the bachelor's degree in computer engineering and the M.Sc. degree in computer systems and networks from the Department of Applied Mathematics, National Technical University of Ukraine "Kyiv Polytechnic Institute," Kyiv, Ukraine, in June 2002 and June 2004, and the Ph.D. degree in computer systems and components from the Department of Information and Computing Systems and Control, West Ukrainian National University, Ternopil, Ukraine, in February 2013. He is an Associate Professor with the Department of Big Data and Artificial Intelligence, School of Computer Science, Hubei University of Technology, Wuhan, China. His research interests include big data analysis, mining technology, intelligent information processing, and visualization technology.



KRZYSZTOF PRZYSTUPA received the master's degree in electrical power engineering from the Lublin University of Technology, Lublin, Poland, in 1994, and the Ph.D. degree in construction and operation of machines from the Lublin University of Technology, in 2000. He is an Assistant Professor with the Department of Automation, Lublin University of Technology. His scientific interests lie in the fields of automation, diagnostics, reliability, quality assurance systems, safety, internet security, instrumentation, and measurements.

OREST KOCHAN received the degree from Physical Faculty, Ivan Franko Lviv National University of Lviv, Lviv, Ukraine, in 2006, and the Ph.D. and D.Eng. degrees in engineering from Lviv Polytechnic National University, Lviv, in 2011 and 2020, respectively. He is with the Department of Information-Measuring Technologies, Lviv Polytechnic National University. His research interests include industrial temperature measurements, modeling in the field of measurements, and the Internet of Things.