

Received September 18, 2021, accepted October 3, 2021, date of publication October 25, 2021, date of current version November 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122543

MH UNet: A Multi-Scale Hierarchical Based Architecture for Medical Image Segmentation

PARVEZ AHMAD¹, HAI JIN¹, (Fellow, IEEE), ROOBAAE ALROOBAAE², SAQIB QAMAR³, RAN ZHENG¹, FADY ALNAJJAR⁴, AND FATHIA ABOUDI⁵

¹National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

²Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia

³Department of Computer Applications, Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh 517325, India

⁴College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates

⁵LRBTM Research Laboratory of Biophysics and Medical Technology, High Institute of Medical Technology in Tunisia (ISTMT), Université Tunis El Manar, Tunis 1006, Tunisia

Corresponding author: Hai Jin (hjin@hust.edu.cn)

This work was supported by the Taif University Researchers Supporting Project, Taif University, Taif, Saudi Arabia, under Grant TURSP-2020/36.

ABSTRACT UNet and its variations achieve state-of-the-art performances in medical image segmentation. In end-to-end learning, the training with high-resolution medical images achieves higher accuracy for medical image segmentation. However, the network depth, a massive number of parameters, and low receptive fields are issues in developing deep architecture. Moreover, the lack of multi-scale contextual information degrades the segmentation performance due to the different sizes and shapes of regions of interest. The extraction and aggregation of multi-scale features play an important role in improving medical image segmentation performance. This paper introduces the MH UNet, a multi-scale hierarchical-based architecture for medical image segmentation that addresses the challenges of heterogeneous organ segmentation. To reduce the training parameters and increase efficient gradient flow, we implement densely connected blocks. Residual-Inception blocks are used to obtain full contextual information. A hierarchical block is introduced between the encoder-decoder for acquiring and merging features to extract multi-scale information in the proposed architecture. We implement and validate our proposed architecture on four challenging MICCAI datasets. Our proposed approach achieves state-of-the-art performance on the BraTS 2018, 2019, and 2020 *Magnetic Resonance Imaging* (MRI) validation datasets. Our approach is 14.05 times lighter than the best method of BraTS 2018. In the meantime, our proposed approach has 2.2 times fewer training parameters than the top 3D approach on the ISLES 2018 *Computed Tomographic Perfusion* (CTP) testing dataset. MH UNet is available at <https://github.com/parvezamu/MHUnet>.

INDEX TERMS BraTS, convolutions, dense connections, encoder-decoder, ISLES, MICCAI, UNet.

I. INTRODUCTION

Convolutional Neural Network (CNN) is a popular approach for medical image segmentation such as brain tumor segmentation [1], stroke lesion segmentation [2], and infant brain tissue segmentation [3]. Manual 3D image segmentation is a time-consuming task due to the variations of each patient's shapes, sizes, and locations. For example, Figure 1 shows three brain MRI patient slices; each tumor has a distinctive shape, size, and location. An accurate automatic

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

segmentation method can speed up clinical decisions in life-threatening problems.

CNN is a sequence of convolution, pooling, non-linearities operations to learn high-level features [4], [5]. However, 2D and 3D CNNs face challenges in obtaining competitive results. For example, Havaei *et al.* [6] and Kamnitsas *et al.* [2] presented multi-scale architectures for local and global features. 2D convolutions in the conventional architecture do not exploit the full contextual information of 3D medical datasets. At the same time, 3D filters provide higher accuracy compared to 2D filters. However, the depth of 3D CNNs is restricted due to the limited resources. Furthermore, the lack

of end-to-end training schemes with traditional 3D CNN degrades the segmentation accuracy.

These limitations can be mitigated by *Fully Convolutional Neural Networks* (FCNNs) such as UNet [7] and its variations [8]–[10]. UNet is made up of an encoder and decoder. The encoder learns the context features and reduces medical images' high resolution by applying convolution and pooling operations. In contrast, the decoder recovers the resolution of medical images by using an upsampling operation. Simultaneously, the decoder adds more abstract representation to the aggregating features of the encoder and the upsampling function by applying convolution operations. In the skip-connection, the aggregation function is either concatenation [8] or addition [9] in the UNet architecture.

Nearly, most CNN methods apply UNet [8], [11], [12] for medical image segmentation. However, these architectures suffer from a huge number of parameters. In addition, several researchers employed cascaded strategies on UNet, especially for brain tumor segmentation [13]–[16] because of overlapped labels. Cascaded UNet involves two or more encoder-decoder networks to solve the problem of segmentation. For example, Baid *et al.* [14] presented a cascaded UNet in which the first encoder-decoder network segments the whole tumor, and the second UNet is trained to segment the tumor core and the enhancing tumor. However, solving a multi-class segmentation problem using the cascading UNet architectures is complicated.

Cascaded UNet also used residual connections [17] to prevent vanishing gradient issues in a deeper network. A deep network allows the encoder-decoder architecture to extract multi-scale contextual information. However, the residual UNet uses several channels during training, which causes an increase in training parameters. Residual network is replaced by dense connections, which allows a sequence of short connections between layers [18]. Dense connections help in reducing parameters and developing a deep network. In a dense network, each layer has a connection with all previous layers. Researchers have developed densely connected UNet for multiple organ segmentation problems [9], [19].

UNet architectures cannot address the issues of heterogeneous organ segmentation, in which regions of interest are inconsistent and vary in size. For example, as seen in Figure 1, the size of brain tumors varies significantly in MRI modalities. Two restrictions are available in the UNet architectures. First, the residual-based UNet architectures take a huge number of parameters for training. Furthermore, in traditional residual blocks, the convolution layers have redundant features. Meanwhile, these layers fail to include previous layers' efficient low-level features in training. Second, current UNet architectures have limited ability to retrieve the contexts of multiple receptive scales efficiently. Researcher presented solutions to overcome this problem [2], [6], [9]. These methods introduced different receptive scales of feature maps. However, these methods cannot deal properly with the diverse medical image modalities that have large-scale variations. Theoretically, the decoder extracts features from deep layers

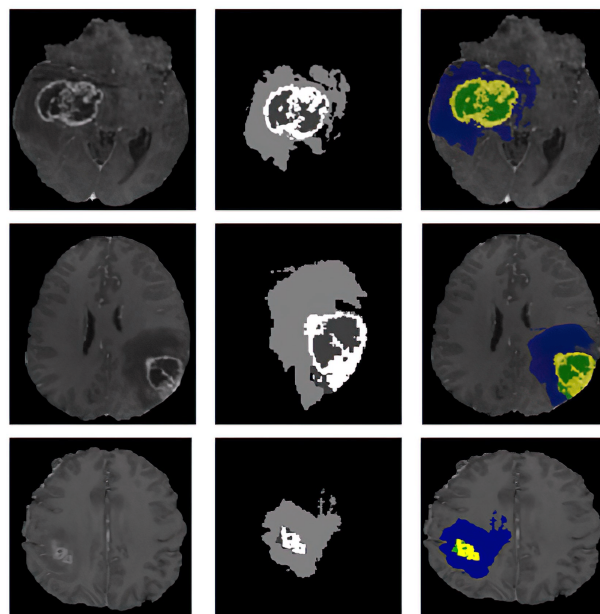


FIGURE 1. MRI scans show the variations of shape, size, and location of brain tumors. Each row depicts T1ce, ground truth and overlaying of ground truth on the T1ce modality of each patient from left to right. Each colour represents a different region: green for the tumor core, blue for the whole tumor, and yellow for enhancing tumor.

in the architecture. Low-resolution features have adequate semantic information, whereas layers of the encoder contain rich spatial information and little global semantic context. Consequently, in the architecture, the high-level semantic and low-level spatial information can be adequately merged to maximize the integration of multi-scale features.

In inspiration of the above concepts, we propose MH UNet to address the issues of heterogeneous organ segmentation. The proposed architecture consists of three blocks. First, we suggest dense blocks in the encoder-decoder to reduce the number of learnable parameters. We replace the residual learning function with a dense connection in each dense block to provide multi-scale features to its adjacent block. As a result, feature maps with different receptive scales are sent into the blocks. Second, we use a residual-inception block comprising three parallel dilated convolution layers to learn local and global contexts in the first level of the encoder. The multi-scale context is then applied to the first dense block of the encoder. As a result, each dense block in the encoder provides multi-scale features to its adjacent block. In the decoder, a variance of the residual-inception block is proposed. As a result, the multi-scale features are available to the dense blocks of the decoder. Third, we offer a unique hierarchical block between the encoder-decoder to efficiently extract features of multiple receptive scales. The encoder's dense blocks extract multi-scale features with fixed receptive scales. Thus, the proposed hierarchical block with dilated convolution layers enhances the sizes of the receptive fields. In particular, we construct a multi-path block of different dilated convolution layers and then combine features

with different paths' receptive fields. As a result, the hierarchical block efficiently fuses high-level semantic features with low-level spatial features to improve the segmentation scores. In addition, a deep supervision concept is presented in the decoder to acquire new semantic features and improve the segmentation maps of multiple organs. The proposed MH UNet outperforms contemporary approaches on four publicly available benchmark datasets.

For segmentation tasks, as a variation of 3D UNet, MH UNet is more profound, flexible, and lightweight. We summarize the key contributions of the proposed work as below:

- We develop a novel multi-scale hierarchical architecture for medical image segmentation. Dense connections allow deep supervision, smoothing the gradients flow, and reduced learnable parameters. Meanwhile, the residual-inception blocks extract multi-scale features for robust representation.
- The hierarchical block efficiently combines the multi-scale local and global contexts in an encoder-decoder architecture. The hierarchical block improves the receptive field sizes of the dense blocks' feature maps by different parallel dilation rates at the encoder of 3D UNet.
- We present a deep supervision approach for faster convergence and superior segmentation accuracy. All dense blocks generate multi-scale segmentation maps in the decoder. These multi-scale segmentation maps are aggregated to boost the model's convergence speed and accuracy.
- We propose a combination of binary cross-entropy and dice loss functions to deal with severe class imbalance problems. Our model achieves significant segmentation accuracy due to the combined loss function, which does not require sophisticated weight hyper-parameter tuning.
- We propose an efficient and simple post-processing technique to eliminate false-positives voxels.
- We have used MICCAI BraTS and ISLES datasets for experimentation. Our proposed model outperformed all other state-of-the-art methods, including cascaded and ensembled approaches.

II. RELATED WORK

This section reviews UNet and its variations, frequently presented for the brain tumor and stroke lesion segmentation tasks.

A. STROKE LESION SEGMENTATION

Both 2D and 3D variations of UNet are extensively available in brain tumor segmentation literature, while 3D UNet requires huge memory requirements, 2D filters ignore the slice level contextual information. The majority of stroke lesion segmentation tasks use 2D UNet and its variations [12], [19]–[22], except some works of 3D UNet [23], [24].

Furthermore, earlier methods used the weight hyper-parameter in the different loss functions to address the data imbalance problem during training. However, predicted results might be biased towards the category with the big weight, which is generally provided to the lesions. Specifically, more false-positive voxels may be associated with predictions. We use a non-weighted loss function for our proposed work to avoid the number of false-positive voxels in the predicted maps.

B. BRAIN TUMOR SEGMENTATION

Cascaded UNet and its variations are used for brain tumor segmentation. Here, different variants of UNet [13]–[16] have solved issues in several stages. Cascaded UNet contains two or three encoder-decoder architectures, which add complexity in solving the segmentation problem. Researchers used single UNet to solve multi-class segmentation problems. Residual based 3D UNet is effective to obtain high accuracy using required depth in a single step. Therefore, researchers used variant forms of 3D Residual UNet [25], [26] to exploit multi-scale contextual information for segmentation. Furthermore, the attention mechanism is used in different variations of UNet [27]–[29] to remove unnecessary and redundant features. However, the variations of UNet are shallow and contain a high number of channels during training, thus require a huge number of training parameters.

In the proposed work, we use densely connected blocks in encoder-decoder. Dense connections reduce the training parameters by using a value of growth-rate in the blocks. We use densely connected residual-inception blocks to extract multi-scale contextual information. A unique hierarchical block between the encoder and decoder is present to extract multi-scale features from numerous receptive scales efficiently. We hierarchically design our mechanism to achieve state-of-the-art performances with a minimum number of parameters.

III. PROPOSED ARCHITECTURE

The proposed architecture for brain tumor segmentation is depicted in Figure 2. The encoder-decoder employ several dense and residual-inception blocks. In addition, we adopted a hierarchical idea inspired by Liu *et al.* [30]. In Figure 17 (see supplementary materials section VII), we visualize MH UNet with channels description. In addition, we also visualize some layers of our proposed work.

A. DENSE BLOCKS

As depicted in Figure 3, a dense block has three convolution layers, in which feature maps of all previous layers are concatenated and passed as input to the current layer. In addition, after the first convolution layer in each dense block, the spatial dropout layer is employed at a rate of 0.2 to avoid overfitting problems. The concept of dense connections in a dense block can be summarized as

$$x_{l+1} = g(x_l) \circ x_l \quad (1)$$

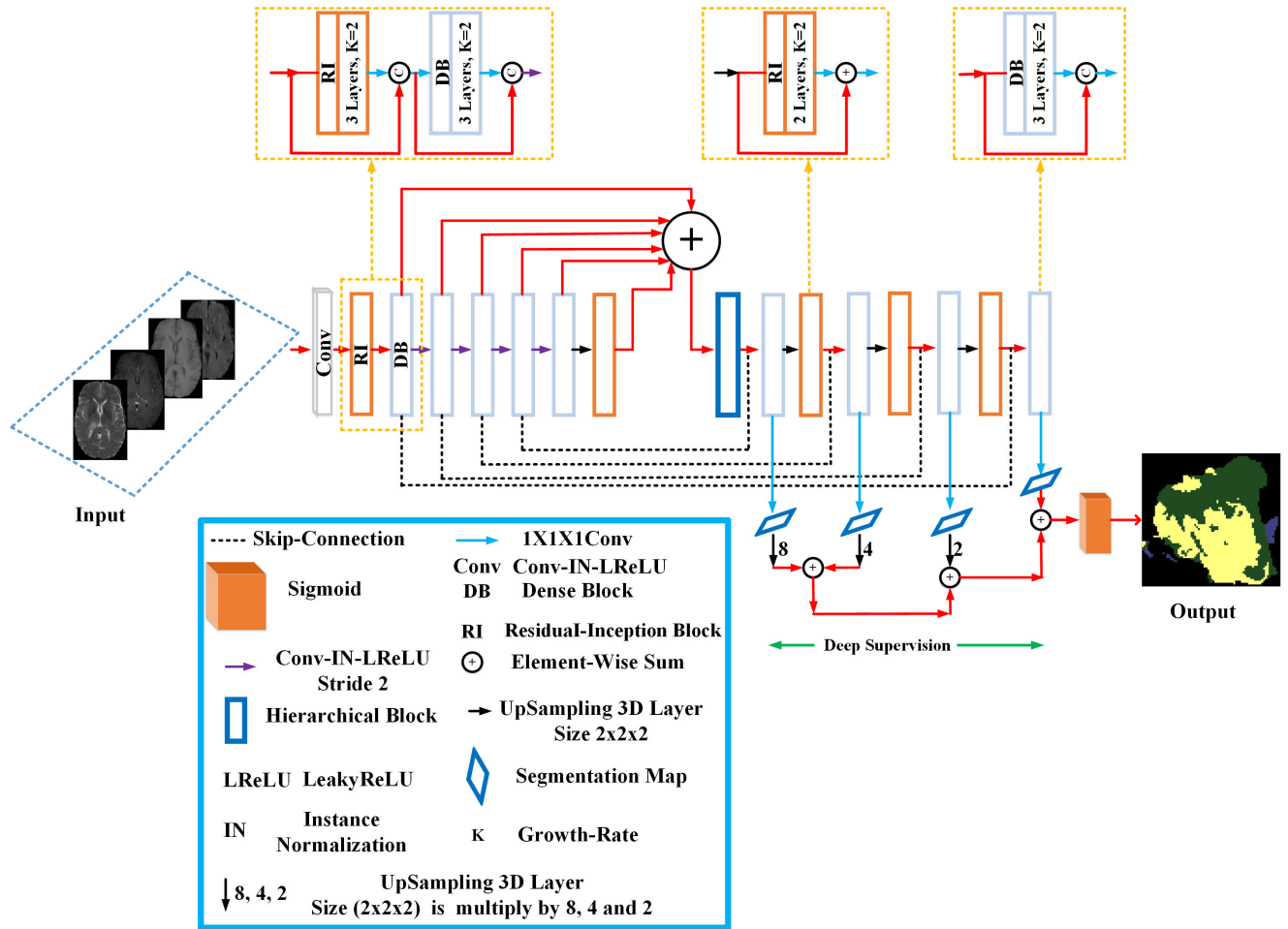


FIGURE 2. Proposed MH UNet. The size of the input to the first convolution layer (Conv) is $128 \times 128 \times 128$. After then, combined feature maps of the residual-inception (RI) block and first convolution layer (Conv) are passed to the dense block (DB). Consequently, the output features have hierarchical information and incorporate the details of all previous layers. The combined information is then transferred via two paths (i) skip-connections, which are used to fuse the features of multiple scales, and (ii) upcoming depths, which are represented by the strided convolution layers (violet arrows) and the dense blocks. Strided convolution layers are employed to reduce the input data’s sizes in the encoder. In contrast, the decoder has non-parametric upsampling layers to retrieve the lost information, followed by the RI blocks for the hybrid contextual learning. In addition, a hierarchical block has aggregated features of the encoder and the first upsampling and RI block. In the meantime, a deep supervision concept is employed on the depths of the decoder for the segmentation maps (blue squares). Finally, an activation function, sigmoid, is applied on the combined segmentation maps for the final output. Figure 3 depicts the details of a dense block, Figure 5 shows variants of residual-inception blocks, and Figure 6 depicts the concept of the hierarchical block.

where x_l denotes the output of a current layer l , $g(\cdot)$ represents a sequence of Conv-IN-LeakyReLU and \oplus denotes a concatenation operation. Furthermore, the input feature maps of a l^{th} convolution layer can be summarized as

$$x_{l+1} = X_{-1} + \sum_{j=1}^l X_j \quad (2)$$

where X_{-1} is the input feature maps for each layer of a dense block.

A growth-rate 2 (X_j) is used in a dense block to reduce the number of parameters in each convolution layer. The input layer uses the feature re-usability property to have more significant features. In this way, dense networks decrease the number of training parameters and the redundant features of a standard 3D convolution. A $1 \times 1 \times 1$ convolution is also

utilized to keep an equal number of input and output channels of a dense block.

The advantages of dense connections in the encoder-decoder are 1) Flow of gradients information easily propagates to all preceding lower layers through short-skip connections. In contrast, the layers without residual and dense connections have an issue of gradients vanishing/exploding. 2) Each dense block offers multi-scale features to its neighbour block. Therefore, feature maps of different receptive scales are input to the blocks. In the proposed architecture, the first dense block has multi-scale inputs resulting from a residual-inception block. 3) Fewer learnable parameters are sufficient for improving the final segmentation scores.

B. RESIDUAL-INCEPTION BLOCKS

Deep learning architectures face challenges of depth and width. In addition, the result of deep learning architectures

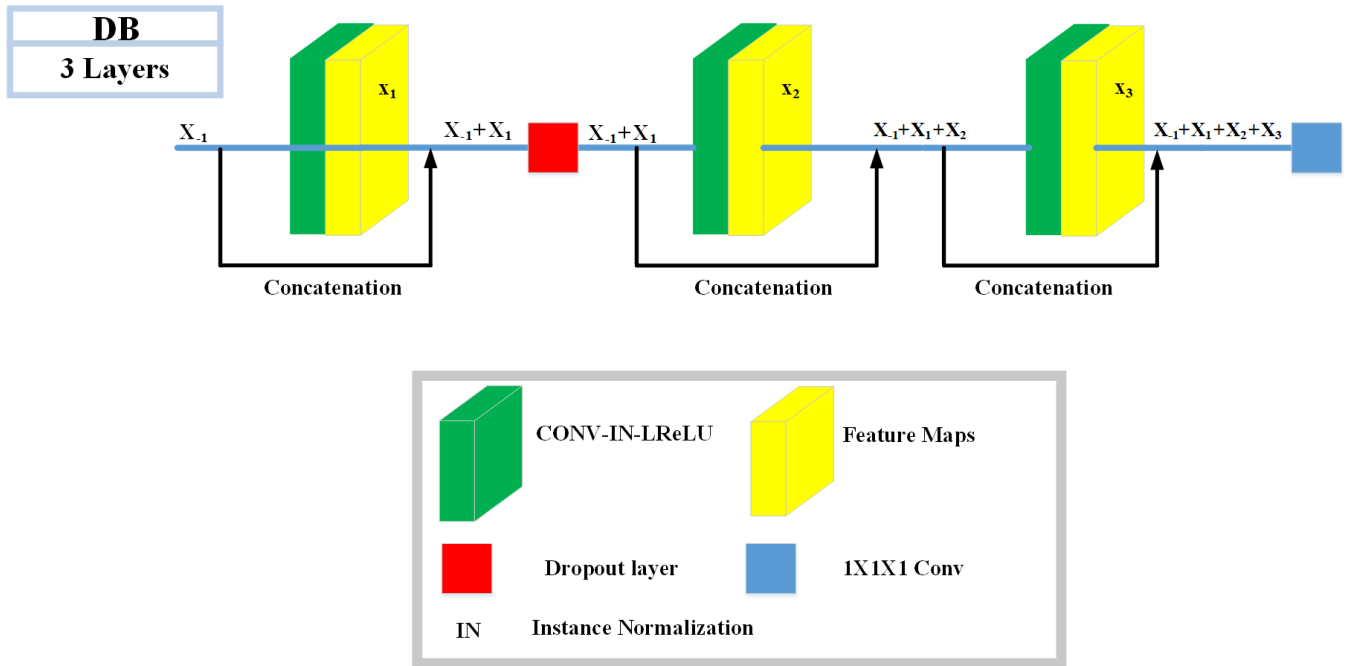


FIGURE 3. Overview of the proposed dense block. X_1 , X_2 , and X_3 denote the feature maps of the convolution layer in terms of growth-rate. Additionally, each convolution layer’s input and output channels are subjected to a concatenation operation.

is degraded by insufficient multi-scale features. Inception blocks [31] with residual connections [32] address the above mentioned issues. We can develop very deep architectures by using inception blocks without increased parameters. Such architecture has sufficient features of multiple receptive scales by the use of the existing inception blocks. Multiple receptive scales generate multi contextual information for segmentation tasks. An inception block, which contains three convolution layers of different receptive field sizes, is depicted in Figure 4a. Moreover, to reduce the number of features, 1×1 convolution layers are used. Finally, aggregating the output features maps of various receptive scales provides multi-scale features. In addition, the residual-based inception block (Figure 4b) solves the issue of the vanishing gradients in the deep architecture.

We apply residual-equipped inception blocks’ multi-scale concepts in our proposed work by substituting 2D layers with 3D layers. Figure 5 illustrates the variants of residual-inception blocks. In the residual-inception block at the encoder (Figure 5a), we employ a dilation rate 2 in the top parallel layer. In contrast, dilation rates 3 and 5 are used in the middle and last parallel layers, respectively. Meanwhile, in the residual-inception blocks of the decoder (Figure 5b), we utilize dilation rates 2 and 3 in the top and last parallel layers, respectively. Different dilation rates increase the receptive field sizes of parallel convolution layers by adding zeros between kernel elements without incrementing parameters. As a result, the proposed residual-inception blocks use large receptive field sizes to learn more local and global contexts. In addition, having multiple dilation rates helps to

avoid the gridding implications that erupt with equal dilation rates [33]. We concatenate input and output feature maps of each receptive scale. We aggregate feature maps convolved by three receptive scales in Figure 5a, such as $5 \times 5 \times 5$, $7 \times 7 \times 7$, and $11 \times 11 \times 11$. Meanwhile, we aggregate feature maps convolved by two receptive scales in Figure 5b, such as $5 \times 5 \times 5$ and $7 \times 7 \times 7$.

Given a convolution layer 1, which has m^l kernel with $m \times m \times m$ size, the effective receptive size or scale of m^l can be described as

$$rf_{m_l} = (r_m - 1) \times (m - 1) + m \quad (3)$$

where r_m is the kernel’s dilation rate and m_l is the size of the kernel. Mathematically, the output of the residual-inception blocks can be expressed as

$$y_{l+1} = ((f_{one} (f_d (y_l) \odot y_l)) \oplus f_{one} (y_l)) \quad (4)$$

where y_l denotes output of a current layer l , $f_d(\cdot)$ represents a sequence of Dilated Conv-IN-LeakyReLU, \odot denotes a concatenation operation, $f_{one}(\cdot)$ represents a sequence of $1 \times 1 \times 1$ Conv-IN-LeakyReLU and \oplus denotes element-wise-sum operation.

C. HIERARCHICAL BLOCK

Because of the substantial size variations in the medical image modalities, the extraction and aggregation of multi-scale features play an important role in improving segmentation precision. UNet and its variations can extract multi-scale features. Such variants, however, are still bound by fixed receptive field restrictions. Furthermore, the dense

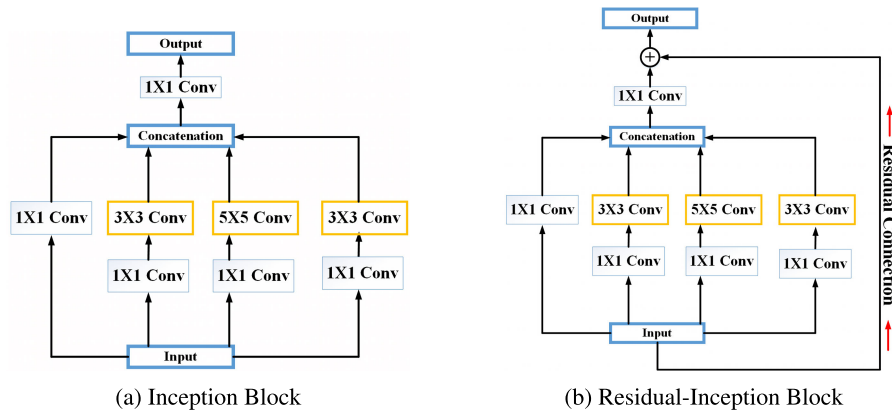


FIGURE 4. Overview of the inception blocks. (a) an inception block without residual connection. (b) a residual-inception block.

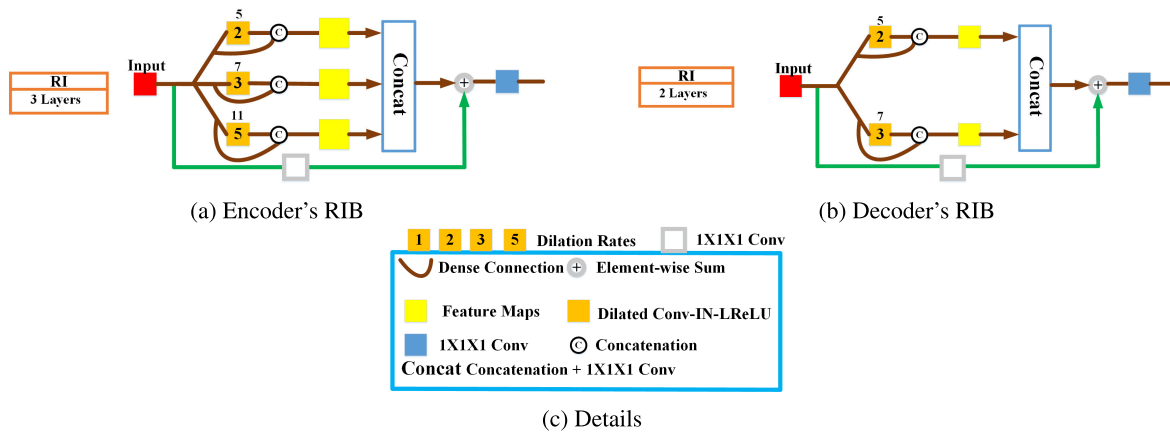


FIGURE 5. Overview of the different residual-inception blocks. (a) a residual-inception block at the encoder (Encoder's RIB). (b) a residual-inception block at the decoder (Decoder's RIB). (c) the details of residual-inception blocks' variants in terms of the layers, residual, and dense connections. IN refers to instance normalization. Dilation rates (orange) 2, 3, and 5 are used in (a). In contrast, for (b), two parallel layers with dilation rates 2 and 3 are used. Meanwhile, the numbers at the top of each dilated convolution layer denote receptive field sizes.

blocks of the encoder produce multi-scale features. However, these features have fixed receptive scales. Therefore, we propose a hierarchical block with dilated convolution layers to increase the receptive field sizes without incrementing the parameters.

In the proposed MH UNet, we present a hierarchical block to extract the multi-scale features of large receptive field sizes to resolve the issue of restricted receptive scales. As shown in Figure 6, we apply the hierarchical block concept by including the output feature maps of all dense blocks of the encoder and the first upsampling layer and residual-inception block of the decoder. The output feature maps of the last dense block at the encoder are fed to the first upsampling layer at the decoder, followed by a residual-inception block to $256 \times 16 \times 16 \times 16$. Both the upsampling layer and the residual-inception block are shown by a single block (orange). Simultaneously, in the encoder, the output feature maps of each dense block are resized to have a tensor shape $256 \times 16 \times 16 \times 16$. To resize them, several downsampling and upsampling operations are used. Finally, the aggregate

feature map (denotes by the element-wise sum operation) is given to the hierarchical block.

As shown in Figure 7, our hierarchical block in the proposed work incorporates feature maps of multi-paths. Each path's feature maps have low-level spatial and high-level semantic details of all numerous dilated convolution layers using the feature re-usability property of dense networks. The feature maps of every two receptive scales of multi-paths are combined into a single feature map. The final step is to concatenate all feature maps from every two receptive scales to show the relevance of feature maps under multiple scales for multi-organ segmentation. The most prominent advantage of the hierarchical block is that it allows for an increase in the receptive field scale of feature maps. A large receptive field scale will be used with multi-scale contexts to identify larger regions of interest. The hierarchical block identifies local regions of interest while crucial multi-scale features are shared amongst the decoder's layers. As shown in Figure 7, the hierarchical block has a component for extracting multi-scale feature information.

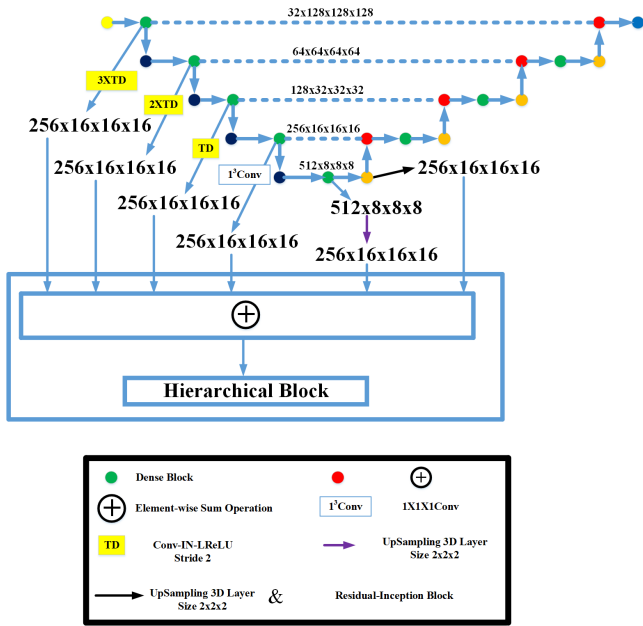


FIGURE 6. Idea of the proposed multi-scale hierarchical block in the MH UNet. The numbers on every skip connection indicate the output feature map of each dense block.

Given an output feature map $FM_{out} \in Q^{OC_{out} \times H \times W \times D}$, which is a combined output of the encoder’s all dense blocks and decoder’s first upsampling layer and a residual-inception

block (see Figure 6), OC_{out} denotes the output channels, the symbols H , W , and D indicate the height, width, and depth of the combined output features map. FM_{out} (denotes by an input predicted image) is initially fed into a $1 \times 1 \times 1$ convolution layer to reduce channel size. To avoid overfitting, the reduced feature maps are fed into a spatial dropout layer (red) with a rate of 0.2. The spatial dropout layer’s features are then shared into two paths. Because of the feature re-usability property of dense networks, each convolution layer with various dilation rates in each path has output feature maps of all previous layers. F_{12} , for example, is the combined output feature map of two separate receptive scales’ feature maps (denotes by F_1 , and F_2 , respectively) that input into a dilated convolution layer (denotes by 2).

$$F_{12} = F_1 \circledast F_2 \tag{5}$$

where \circledast denotes the concatenation operation. Consider another output feature map, F_{1123} , which feeds into the final dilated convolution layer (denotes by 3) and contains output feature maps from the first part’s preceding layers.

$$F_{1123} = F_1 \circledast F_{12} \circledast F_3 \tag{6}$$

In the second part of the proposed hierarchical block, we can observe similar output feature maps that feed into several dilated convolution layers. The output feature maps

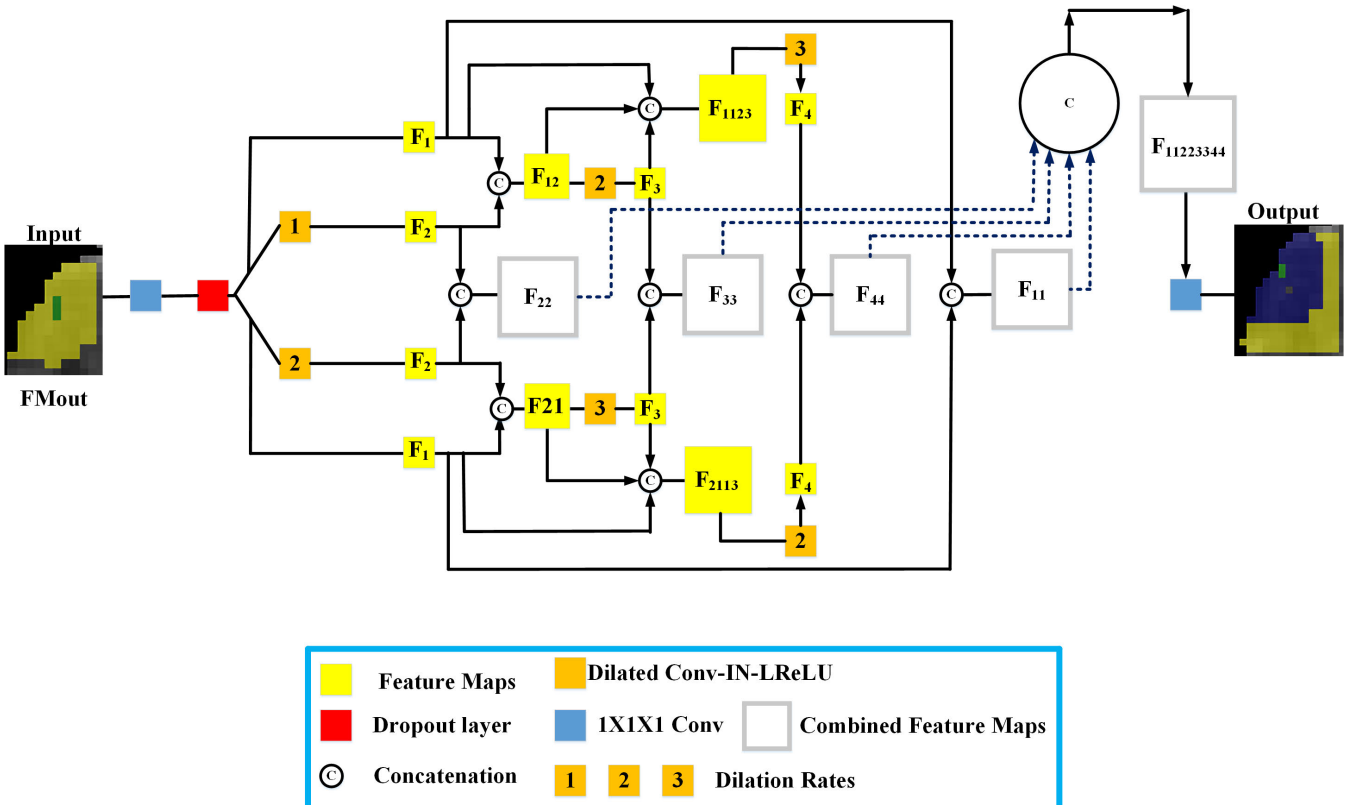


FIGURE 7. Overview of the hierarchical block. Each colour represents a different tumor: green for the tumor core, blue for the whole tumor, and yellow for enhancing tumor.

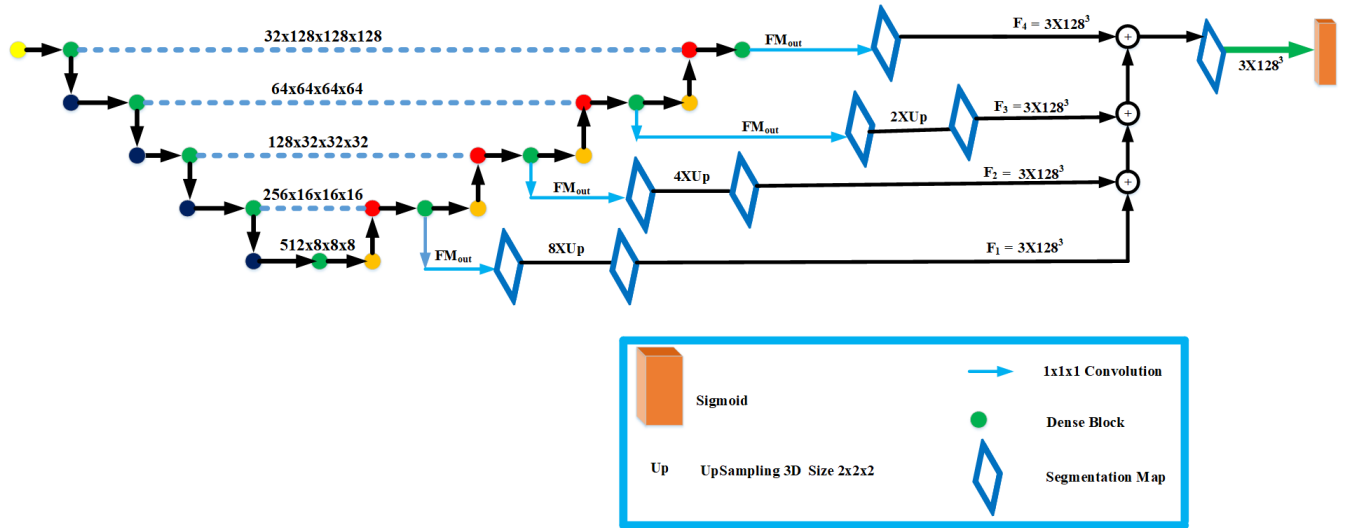


FIGURE 8. Overview of the deep supervision mechanism.

of each two parallel dilated convolution layers are then concatenated (denotes by F_{22} , F_{33} , F_{44} , respectively) to extract more efficient multi-scale contexts.

$$F_{22} = F_2 \circledast F_2 \quad (7)$$

$$F_{33} = F_3 \circledast F_3 \quad (8)$$

$$F_{44} = F_4 \circledast F_4 \quad (9)$$

Finally, all parallel dilated convolution layers' output feature maps, including the spatial dropout layers (denotes by F_{11}), are concatenated (denotes by $F_{11223344}$) to include features from all receptive scales.

$$F_{11223344} = F_{11} \circledast F_{22} \circledast F_{33} \circledast F_{44} \quad (10)$$

This combined output feature map is then fed into a $1 \times 1 \times 1$ convolution layer for feature reduction (denotes by an output predicted image) before the first fusion operation (see Figure 6) to equal size output feature maps is performed. The multi-scale features are further enhanced in the higher layers of the proposed MH UNet.

D. DEEP SUPERVISION

In the decoder, we also propose a deep supervision technique [34] on the output feature maps of several dense blocks for faster convergence and superior segmentation accuracy. To reduce the tensors' features, all dense blocks' output feature maps are fed into a $1 \times 1 \times 1$ convolution layer, as shown in Figure 8. The result is represented as segmentation maps (denoted by the blue squares). Meanwhile, we use upsampling layers of varying sizes (denotes by 8, 4, and 2) to make the size of segmentation maps equal to the input patch size. The updated segmentation maps are then subjected to an element-wise summation. As a result, the proposed deep supervision strategy allows for more direct backpropagation

to the deep layers, potentially avoiding optimization issues. Finally, we apply a sigmoid activation function to all aggregate segmentation maps to densify the classification output. As a result, in addition to the final layer's (denotes by the tensor of shape $32 \times 128 \times 128 \times 128$) segmentation map, the proposed architecture has three more same-resolution segmentation maps to improve the final segmentation results.

Consider the output feature maps of dense blocks as $FM_{out} \in Q^{OC_{out} \times H \times W \times D}$, OC_{out} denotes the output channels, the symbols H , W , and D indicate the height, width, and depth of the dense blocks' output features maps. FM_{out} are then fed into a $1 \times 1 \times 1$ convolution layers to reduce the tensors' features. The resulting features (shown by a segmentation map in blue squares) are then fed into upsampling layers, which use various sizes to increase the size of reduced features to equal the network's input resolution. As a result, the segmentation maps F_1, F_2, F_3, F_4 can be summarised as follows:

$$F_1 = \text{Upsampling3D}_8 \left(f_{one} \left(y_{l_{decoder}_{128 \times 16^3}} \right) \right) \quad (11)$$

$$F_2 = \text{Upsampling3D}_4 \left(f_{one} \left(y_{l_{decoder}_{64 \times 32^3}} \right) \right) \quad (12)$$

$$F_3 = \text{Upsampling3D}_2 \left(f_{one} \left(y_{l_{decoder}_{32 \times 64^3}} \right) \right) \quad (13)$$

$$F_4 = f_{one} \left(y_{l_{decoder}_{16 \times 128^3}} \right) \quad (14)$$

where 8, 4, and 2 are upsample sizes of Upsampling3D layers. FM_{out} stands for 128×16^3 , 64×32^3 , 32×64^3 , 16×128^3 and refers to the dense blocks' output feature maps. $f_{one}(\cdot)$ denotes a sequence of $1 \times 1 \times 1$ Conv-IN-LeakyReLU. Finally, to avoid optimization complications and improve segmentation accuracy, we perform element-wise addition operations between each pair of segmentation maps.

IV. MICCAI BraTS CHALLENGES

A. DATASETS

The BraTS datasets for the years 2018, 2019, and 2020 [35], [36] were collected from 19 different medical institutions. These institutions used a variety of MRI scanners and imaging procedures to obtain MRI scans. All of the brain scans are co-registered, skull-stripped, and interpolated in the meantime. The dimension of each MRI modality is $240 \times 240 \times 155$. Each BraTS dataset (2018, 2019, and 2020) is divided into two parts: training and validation. *High-grade glioma* (HGG) and *low-grade glioma* (LGG) are the two types of glioblastoma covered in the training part of BraTS 2018, 2019, and 2020. The HGG part of the BraTS 2018 training dataset comprises 210 MRI patients, whereas the HGG parts of the BraTS 2019 and 2020 training datasets have 259 and 293 patients, respectively. Meanwhile, the LGG part of the BraTS 2018 training dataset has 75 patients, and the LGG part of the BraTS 2019, and 2020 training datasets has 76 patients. The validation part of the BraTS 2018 dataset comprises 66 patients, whereas the BraTS 2019, and 2020 validation parts have 125 patients. The four MRI modalities are T1, T1ce, T2, and FLAIR for each patient. The training datasets include a truth-label for each patient. Each predicted MRI has three tumors: *whole tumor* or *whole* or *WT*, *tumor core* or *core* or *TC*, and *enhancing tumor* or *enhancing* or *ET*. Labels 1, 2, and 4 are used to evaluate a whole tumor. Labels 1 and 4 are combined to determine the tumor core. Label 4 is used to determine whether a tumor is enhancing. The truth-label is not provided for the BraTS validation datasets' patients.

B. IMPLEMENTATION DETAILS

The training of our proposed model starts after normalizing each MRI modality. After normalization, patches of size $128 \times 128 \times 128$ are created from the BraTS training datasets. During the training, we use Adam optimizer and set 1 as the batch size. The initial learning rate during training is set to 4×10^{-5} , but if validation loss is not improved within 20 epochs, the rate is reduced by a factor of 0.5. The network is trained for 300 epochs. Additionally, augmentation techniques such as random rotation and flipping are used to avoid the overfitting problem during the training.

During the training of the network, the following combined loss function is used

$$L^{totalloss} = -(L_{MDL} - L_{BCL}) \quad (15)$$

where MDL is the multi-label dice loss function [8] and BCL represents the binary cross-entropy loss. Mathematically, MDL , and BCL can be written as

$$Loss^{MDL} = \frac{2}{D} \sum_{d \in D} \frac{\sum pred_d truth_d}{\sum pred_d + \sum truth_d} \quad (16)$$

$$Loss^{BCL} = -\frac{1}{T} \sum_{d \in D} \sum (truth_d \cdot \log(pred_d)) + (1 - truth_d) \cdot \log(1 - pred_d) \quad (17)$$

where $pred_d$ and $truth_d$ is the prediction and ground truth of class d , respectively. D is the total number of classes. T is the number of voxels in output.

The proposed model is implemented in Keras, with 32 GB GPU memory.

C. EVALUATION

We evaluate the predictions of the BraTS datasets. Several metrics, such as the *Dice Similarity Coefficient* (DSC), the Sensitivity, the Specificity, and the Hausdroff95 distances, are used to evaluate the predicted labels. In the BraTS public benchmark dataset [36], each of the metrics is very thoroughly described. In all prior BraTS challenges, DSC was the deciding metric for winning teams; thus, we report the best method based on the highest average DSC or dice scores.

D. POST-PROCESSING

There are no enhancing tumors in LGG cases of the BraTS datasets. The absence of an enhancing tumor will have a significant impact on overall segmentation accuracy. The issue arose when small blood regions in the tumor core failed to be predicted as necrosis or as the whole tumor. This concern raises the issue of false-positives for the enhancing tumor. To eliminate false-positives, we employ a post-processing technique of this study [10]. For the BraTS 2018 and 2019 datasets, all predicted enhancing tumor regions are replaced with necrosis when the threshold value is less than 450 voxels. This reduces the number of false-positive voxels from the predicted training and validation sets. Meanwhile, for the BraTS 2020 datasets, a threshold value less than 575 voxels works well.

The algorithmic details of the post-processing step are shown in Algorithm 1. Here, the enhancing tumor voxel (V_{ET}) of each predicted MRI (PS_i) is extracted (E). The next step is to count (C) all tumor voxels that are enhancing. If the total sum of enhancing tumor regions (TS_{ET}) is less than 450 or 575 voxels (Th), the necrosis replaces (RW_{TC}) all of the predicted enhancing regions. After the post-processing step, the predicted MRI ($\sum_{i=1}^n PP_i$) is aggregated and made available as post-process predicted MRIs (PP).

Algorithm 1 Post-Processing Step

Input: Proposed architecture predicted MRI (PS)

Output: Post-process predicted MRIs (PP)

```

1: for  $PS_i \in PS$  do
2:    $V_{ET} \leftarrow E_{ET}(PS_i)$ 
3:    $TS_{ET} \leftarrow C(V_{ET})$ 
4:   if  $(TS_{ET} < Th)$  then
5:      $RW_{TC} \leftarrow TS_{ET}(PS_i)$ 
6:      $PP_i \leftarrow RW_{TC}$ 
7:   end if
8: end for
9:  $PP \leftarrow \sum_{i=1}^n PP_i$ 
10: return PP

```

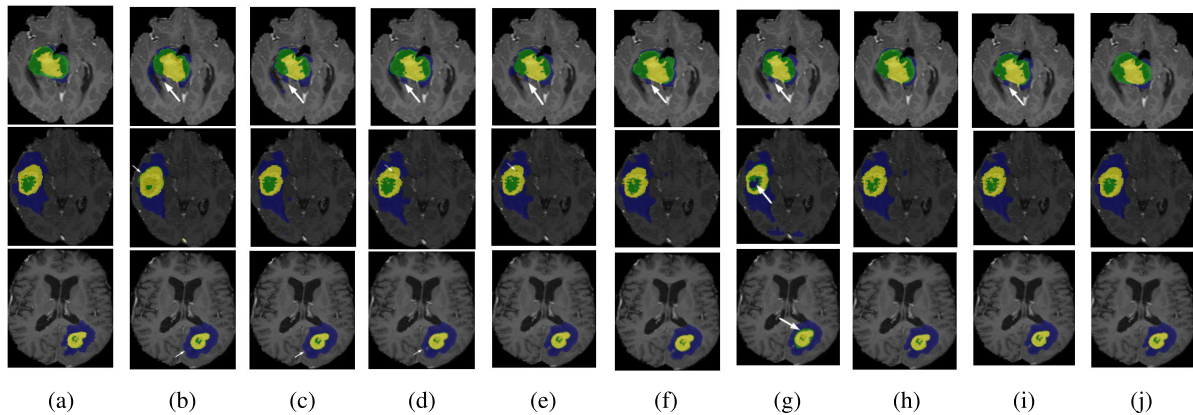


FIGURE 9. Segmentation results on the BraTS 2018 training dataset. From (a) to (j): ground-truth, B + R, B + D + DS, B + D + RI_Wo_DR + DS, B + D + RI_W_DR_V1 + DS, B + D + RI_W_DR_V2 + DS, B + D + H_Wo_DR + DS, B + D + H_W_DR + DS, B + D + RI + H + Wo_DS, and B + D + RI + H + W_DS results overlaid on T1ce modality; whole tumor (blue), tumor core (green), and enhancing tumor (yellow). A white arrow denotes the mis-segmentations.

TABLE 1. Statistical analysis of ablation studies on the BraTS 2018 validation dataset. Each row is a method, where B denoted a baseline (MH UNet without dense (D), residual-inception (RI), and hierarchical (H) blocks, and deep supervision mechanism), and R denoted the residual blocks. Meanwhile, row numbers 4 to 5 denoted the multiple variants (denoted by V1 and V2) of dilation rates for RI blocks. In each variant, the first bracket of dilation rate/s belongs to the encoder, while the second one indicates the dilation rates of the decoder’s RI blocks. The best average dice score of three tumors (denoted by average) is highlighted in bold.

Methods	R	D	RI DR	H DR	DS	Average	Parameters
B + R	✓	✗	✗	✗	✗	81.594	8.2 M
B + D + DS	✗	✓	✗	✗	✓	83.784	2.9 M
B + D + RI_Wo_DR + DS	✗	✓	✓(1), (1)	✗	✓	82.360	3.1 M
B + D + RI_W_DR_V1 + DS	✗	✓	✓(1, 2, 3), (1,2)	✗	✓	83.951	3.1 M
B + D + RI_W_DR_V2 + DS	✗	✓	✓(2, 3, 5), (2,3)	✗	✓	84.916	3.1 M
B + D + H_Wo_DR + DS	✗	✓	✗	✓(1)	✓	82.442	3.5 M
B + D + H_W_DR + DS	✗	✓	✗	✓(1, 2, 3)	✓	83.500	3.5 M
B + D + RI + H + Wo_DS	✗	✓	✓(2, 3, 5), (2,3)	✓(1, 2, 3)	✗	85.070	3.7 M
B + D + RI + H + W_DS	✗	✓	✓(2, 3, 5), (2,3)	✓(1, 2, 3)	✓	85.400	3.7 M

E. ABLATION STUDIES

We run various experiments to evaluate the efficacy of the blocks that make up the MH UNet. After removing the proposed dense (denoted by D), residual-inception (denoted by RI), hierarchical (denoted by H) blocks, and deep supervision (denoted by DS), we use the resulting MH UNet as a baseline (denoted by B). The visual and dice score comparisons for several blocks and DS are shown in Figure 9 and Table 1, respectively. The dense blocks, as described in sub-section III-A, reduce the number of learnable parameters. As a result, we test the efficacy of D connections (denoted by “B + D + DS”) in lowering the learnable parameters. In the meantime, enormous learnable parameters can be observed when we use residual blocks with the baseline (denoted by “B + R”). The RI and H blocks extract multi-scale features, as indicated in sub-sections III-B and III-C. As a result, we run additional tests using standard (denoted by “B + D + RI_Wo_DR + DS”) and “B + D + H_Wo_DR + DS”) and dilated convolutions (denoted by “B + D + RI_W_DR_V1 + DS”) and “B + D + RI_W_DR_V2 + DS”) and “B + D + H_W_DR + DS”) to validate the efficacy of the multi-scale features. We also conduct tests (denoted by “B + D + RI_W_DR_V1 + DS”, and “B + D + RI_W_DR_V2 + DS”) to determine

the best dilation rates (denoted by DR) for RI blocks’ variants. Finally, we conduct two trials (denoted by “B + D + RI + H + Wo_DS”) and “B + D + RI + H + W_DS”) to validate the efficacy of deep supervision (denoted by DS).

First, we apply the proposed dense blocks and deep supervision technique (denoted by “B + D + DS”) to the BraTS 2018 dataset. Figure 9 shows two typical brain tumor segmentation results (see Figure 9b and Figure 9c), demonstrating that our suggested dense blocks may successfully reduce mis-segmentation outcomes that the baseline cannot manage well with residual blocks (denoted by “B + R”). As demonstrated in Table 1, “B + D + DS” improves the average performance of the DSC metric from 81.594% to 83.784% compared to “B + R”. Furthermore, dense blocks lower the number of parameters from 8.2 M to 2.9 M when compared to residual blocks.

Second, we investigate the effectiveness of the RI blocks. As shown in Table 1, compared with “B + D + RI + W_DR_V1 + DS”) and “B + D + RI + W_DR_V2 + DS”, we can see that the average score of the DSC metric of “B + D + RI + Wo_DR + DS”) decreases first from 82.360% to 83.951% and then from 82.360% to 84.916%, demonstrating that the multi-scale features can improve

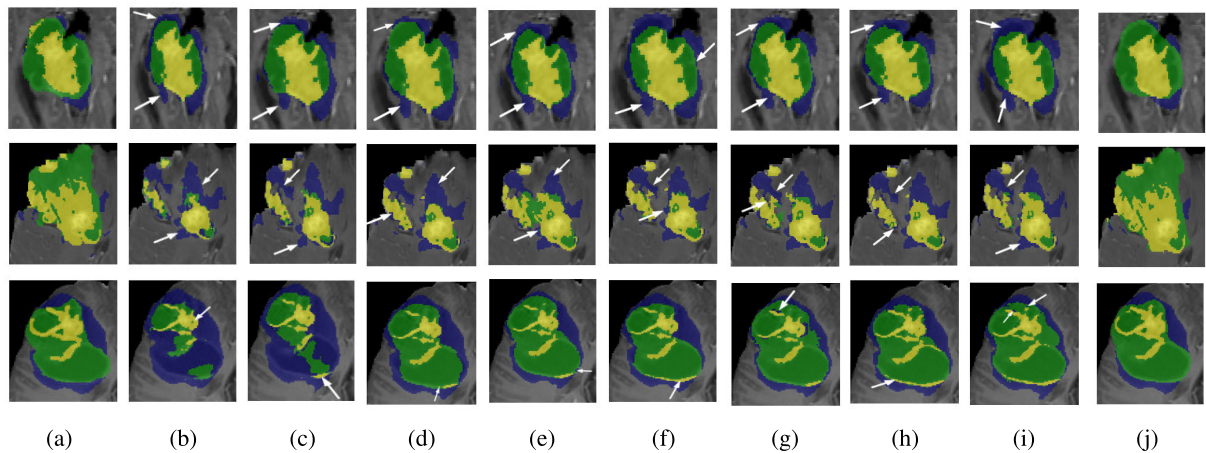


FIGURE 10. Segmentation results on the BraTS 2018 training dataset. From (a) to (j): ground-truth, Zhao *et al.* [BL + warmup + fuse + psudo label] [37], McKinley *et al.* [Filtered Output] [38], Lachinov *et al.* [Res UNet] [39], Wang *et al.* [TransBTS] [11], Isensee *et al.* [Ensemble] [10], Vu *et al.* [Ensemble (7 models)] [40], Fidon *et al.* [Ensemble] [41], Ghaffari *et al.* [Ensemble] [13], and MH UNet results overlaid on T1ce modality; whole tumor (blue), tumor core (green), and enhancing tumor (yellow). A white arrow denotes the mis-segmentations.

segmentation accuracy. The effectiveness of multi-scale features can be further justified with mis-segmentation outcomes, which are maximum with non-dilation RI blocks (Figure 9d) and minimum with dilated RI blocks (see Figure 9e and Figure 9f). Compared with “B + D + RI + W_DR_V1 + DS”, we can see that the average score is 84.916% when we employ dilation rates 2, 3, and 5 in encoder’s RIB and dilation rates 2, and 3 in decoder’s RIBs (denoted by “B + D + RI + W_DR_V2 + DS”). Predictions are almost similar to ground truth, employing large dilation rates in the RIBs of encoder-decoder, as shown in Figure 9f.

Third, we investigate the effectiveness of the hierarchical (denoted by H) block. As shown in Table 1, compared with the “B + D + H_Wo_DR + DS”, the proposed H block (denoted by “B + D + H_W_DR + DS”) increases the average DSC scores by 10.6% (from 82.442% to 83.500%). Figure 9h shows a typical example of a brain tumor segmentation result, which demonstrates that the multi-scale features of proposed H block can effectively minimize the mis-segmentation outcomes, which the fixed receptive scales cannot well handle (denoted by “B + D + H_Wo_DR + DS”).

MH UNet’s deep supervision mechanism (denoted by DS) is also examined. “B + D + RI + H + W_DS” achieves the highest average DSC scores when compared to “B + D + RI + H + Wo_DS”, even without increasing the extra parameters. The multi-scale outputs of the DS allow “B + D + RI + H + W_DS” (our model) achieve more accurate segmentation results than “B + D + RI + H + Wo_DS” as shown in Figure 9j. As shown in Table 1, our model improves average DSC scores by 38% compared to “B + R”. By combining D, RI, and H blocks and deep supervision (denoted by DS) in a seamless manner, we can see that our method makes tremendous improvements. This indicates the efficacy of MH UNet in dealing with the large-scale variations of brain tumors.

TABLE 2. The comparison of different approaches’ performances on the BraTS 2018 training dataset. The dice scores are reported as a mean. The best value of the dice scores is highlighted in bold.

Methods	Dice Scores		
	WT	TC	ET
Zhao <i>et al.</i> [BL + warmup + fuse + psudo label] [37]	91.400	85.600	78.000
McKinley <i>et al.</i> [Filtered Output] [38]	92.701	88.105	80.326
Lachinov <i>et al.</i> [Res UNet] [39]	92.150	89.350	75.690
Wang <i>et al.</i> [TransBTS] [11]	90.880	90.012	81.980
Isensee <i>et al.</i> [Ensemble] [10]	91.330	84.660	77.890
Vu <i>et al.</i> [Ensemble (7 models)] [40]	93.950	91.265	80.745
Fidon <i>et al.</i> [Ensemble] [41]	93.001	87.770	79.950
Ghaffari <i>et al.</i> [Ensemble] [13]	91.000	83.000	79.000
(Proposed)	93.319	91.469	82.130

F. COMPARISON WITH THE BASELINE APPROACHES

In this sub-section, our proposed work is compared with the following baseline approaches: Zhao *et al.* [BL + warmup + fuse + psudo label] [37], McKinley *et al.* [Filtered Output] [38], Lachinov *et al.* [Res UNet] [39], Wang *et al.* [TransBTS] [11], Isensee *et al.* [Ensemble] [10], Vu *et al.* [Ensemble (7 models)] [40], Fidon *et al.* [Ensemble] [41], and Ghaffari *et al.* [Ensemble] [13]. These approaches are discussed in the subsequent sub-sections.

These approaches, including our proposed MH UNet, use the BraTS 2018 training dataset for training. For each approach, 228 patients are used for training, and the remaining patients are used for validation. The average DSC score of each approach is shown in Table 2. Our proposed approach secures the best mean dice scores for the tumor core and enhancing tumor as shown in Table 2. Meanwhile, the results show that the proposed work obtains a lower whole tumor score than Vu *et al.* [Ensemble (7 models)] [40].

Figure 10 shows a comparison between the predictions of different approaches. The predictions of multiple approaches are overlaid on T1ce modalities. The lack of multi-scale features increases mis-segmentations (denoted by a white arrow) on baseline approaches. Meanwhile, the proposed work exactly matches the ground-truth.

TABLE 3. The comparison of different approaches' performances on the BraTS 2019 validation dataset. The metrics' scores are reported as a mean. The best value of the scores is highlighted in bold.

Methods	Dice Scores			Hausdorff95(mm)		
	WT	TC	ET	WT	TC	ET
Zhao [BL] [37]	89.300	80.800	70.200	5.078	6.472	4.766
Zhao [BL + warmup] [37]	90.400	80.200	72.900	4.141	8.099	3.832
Zhao et al. [BL + warmup + fuse] [37]	90.800	82.300	73.700	4.599	6.433	4.089
Zhao et al. [BL + warmup + fuse + psudo label] [37]	91.000	83.500	75.400	4.569	5.581	3.844
McKinley et al. [RawOutput] [38]	91.000	81.000	71.000	5.970	6.210	4.350
McKinley et al. [Filtered Output] [38]	91.000	83.000	77.000	4.520	6.270	3.920
Lachinov et al. [UNet] [39]	89.740	83.490	74.020	–	–	–
Lachinov et al. [Res UNet] [39]	90.180	82.780	74.240	–	–	–
S. Wang et al. [Ensemble] [28]	90.000	81.000	75.000	4.700	7.110	4.990
Chen et al. [AMPNet][32 Channels] [42]	89.340	79.480	74.130	6.029	7.926	5.270
Chen et al. [AMPNet + TTA][32 Channels] [42]	90.260	79.250	74.160	4.378	7.954	4.575
Wang et al. [TransBTS] [11]	88.890	81.410	78.360	7.599	7.584	5.908
Vu et al. [Ensemble (9 Models)] [15]	90.340	81.120	78.420	4.320	6.280	3.700
(Proposed)	90.262	83.469	78.561	5.689	7.093	3.591

G. RESULTS OF MICCAI BraTS CHALLENGES

1) MICCAI BRATS 2019 CHALLENGE

A comparison of the proposed model and state-of-the-art techniques is shown in Table 3. We find that the dice score of the enhancing tumor with the presented work is superior to that of other methods, including the top-ranked approaches (denotes by first six rows of Table 3). Meanwhile, the average dice score of the whole tumor is lower than the top techniques [37], [38]. At the same time, the tumor core's mean dice score is nearly identical to the method of Lachinov et al. [UNet] [39].

Zhao et al. [BL + warmup + fuse + psudo label] [37] used a range of patch strategies, network ensembling architectures, and learning design approaches to achieve the best score of the whole tumor. In all forms of Zhao et al. [37], large input patch sizes have improved contextual information. Zhao et al. [BL] [37] uses an ensemble technique to analyze the five predicted sets of 3D UNet. Furthermore, Zhao et al. [BL + warmup] [37] has claimed the best score of the whole tumor, which is enhanced further by performing an ensembling technique on six different 3D UNet variations, as reported in Zhao et al. [BL + warmup + fuse] [37]. In contrast, our single proposed work, which did not use a complex ensembling strategy, has a lower score for the whole tumor while achieving the highest scores for the tumor core and enhancing tumor. At the same time, McKinley et al. [RawOutput] [38] used a weight ensemble approach to obtain the optimal value for the whole tumor. For individual directions of the 3D MRI, a shallow variation of 2D UNet with multiple dense blocks is proposed. An attention mechanism has been used between the dense blocks to retain the important features. Finally, on combining the multiple directions' 2D UNets, the best score of the whole tumor is reported. Furthermore, using an uncertainty filtering strategy to ensemble predictions, McKinley et al. [FilteredOutput] [38] reports improved

scores for the tumor core and enhancing tumor. Compared to multiple McKinley [38] variants, our single proposed method receives the highest values for the tumor core and enhancing tumor. Lachinov et al. [39] proposes two distinct 3D UNet variations. The first form of UNet, Lachinov et al. [UNet] [39], uses 32 initial channels for the best score of the tumor core. In contrast, we achieve an almost similar score for the tumor core using exactly half of the initial channels. As a result, our proposed method yields the highest scores for the whole tumor and the enhancing tumor with fewer parameters. In addition, Lachinov et al. [Res UNet] [39] has been reported as a 3D Unet variation. For better feature use, this variation makes use of residual connections. However, the residual-based variation has lower scores for the tumor core with the largest input patch sizes. Wang et al. [Ensemble] [28] revealed the final scores of all tumors after merging multiple 3D UNet variations. Multiple loss functions, including weight hyper-parameter tuning-based losses, were used to train these variations. In addition, an attention technique was used to extract only the most significant features. Our single proposed deepest approach, which is trained without any weight hyper-parameter tuning, has improved all tumors' final scores. Chen et al. [42] proposed two 3D UNet variants. Chen et al. [AMPNet][32 Channels] [42] uses 32 initial channels to train a deep supervision-based 3D variant. However, because of the limited depth, the reported scores are lower. Even with a test-time augmentation strategy, the shallow depth of the 3D UNet variation (Chen et al. [AMPNet + TTA][32 Channels] [42]) has little effect on the final scores. To improve the multi-scale features of current UNet variations, Wang et al. [TransBTS] [11] presented a transformer approach using 3D UNet. The transformer technique is used to improve the contextual information of the higher layers. The current transformer methodology, however, is trained on many GPUs. The transformer approach's

TABLE 4. The comparison of different approaches' performances on the BraTS 2018 validation dataset. The metrics' scores are reported as a mean. The best value of the scores is highlighted in bold.

Methods	Dice Scores			Hausdorff95(mm)		
	WT	TC	ET	WT	TC	ET
Isensee <i>et al.</i> [Ensemble] [10]	91.260	86.340	80.660	4.270	6.520	2.410
McKinley <i>et al.</i> [43]	90.300	84.700	79.600	4.170	4.930	3.550
Zhou <i>et al.</i> [27]	90.780	85.750	81.111	4.884	6.932	2.881
Kermi <i>et al.</i> [44]	86.800	80.500	78.300	8.127	9.840	3.728
Albiol <i>et al.</i> [VGG-like] [45]	87.200	76.000	75.100	-	-	-
Albiol <i>et al.</i> [Inception2] [45]	87.700	77.300	75.330	-	-	-
Albiol <i>et al.</i> [Inception3] [45]	87.300	77.600	78.100	-	-	-
Albiol <i>et al.</i> [Densely Connected] [45]	87.400	75.500	72.900	-	-	-
Albiol <i>et al.</i> [Ensemble] [45]	88.100	77.700	77.300	-	-	-
Feng <i>et al.</i> [Ensemble] [46]	90.940	83.620	79.170	3.801	5.645	4.018
(Proposed)	90.715	84.115	81.374	4.202	7.534	2.803

potential for medical image segmentation is limited due to its high processing resource requirements. Our proposed work, on the other hand, is trained on a single GPU. Furthermore, the best results from our proposed work have shown the efficacy of superior multi-scale features. Vu *et al.* [Ensemble (9 Models)] [15] proposed a cascaded approach of 3D UNets. Residual learning and squeeze and excitation concepts are all incorporated into each 3D UNet. Several 3D UNets are used in the current cascaded technique. The number of 3D UNets is the same as the number of available labels. That is, each 3D UNet is utilized to segment individual labels. Three 3D UNets, for example, are used for three labels of BraTS datasets. The final results are computed by merging the individual segmented labels of 3D UNets. As a result, the 3D UNets' cascaded technique increases the difficulty of solving the segmentation problem. In addition, nine of these cascaded approaches are subjected to an ensemble approach. When compared to existing techniques, our proposed work has the best mean value of the enhancing tumor. It also provides competitive mean scores for the whole tumor and the tumor core without complex ensemble approaches, weight-loss functions, or bigger patch sizes. Furthermore, all of the techniques in Table 3 used either the attention mechanism or the ensembling operation to average training and validation predictions or a combination of both attention and ensembling operations with encoder-decoder architectures. Furthermore, these techniques learn a high number of parameters since their initial learning channels are large. We only consider DSC metric scores because they were the top model's only criterion in previous BraTS competitions. Nonetheless, for the Hausdorff95 distances, our proposed technique provides the best mean score of the enhancing tumor.

2) MICCAI BRATS 2018 CHALLENGE

In terms of metrics, DSC, and Hausdorff95 distances, Table 4 presents a comparison between the proposed model and

state-of-the-art approaches. Only the mean DSC scores are taken into account for the best techniques. In the ensemble methodology of Isensee *et al.* [10], we can see the best mean dice scores of the whole tumor and tumor core in Table 4. Meanwhile, compared to all ensemble approaches like [10], [27], [43], [46], our single model has the best mean score of the enhancing tumor.

Isensee *et al.* [10] provided the final scores using a variation of the original 3D UNet model after conducting the ensemble and a post-processing step. The aggregate weights of multiple trained 3D UNet variations are employed to evaluate the various BraTS 2018 datasets. Finally, false-positive voxels in predicted MRIs are removed using a post-processing approach. McKinley [43] proposed a *CNN* ensemble method. For individual directions of the 3D MRI, two shallow variations of 2D UNet that utilize numerous dense blocks are proposed. The dense blocks contain numerous densely connected convolution layers with varying dilation rates to expand the receptive field sizes. Finally, the improved score of the tumor core is presented when the three directions' shallow 2D UNets are merged. While both approaches (denotes by the top two rows of Table 4) produce the best scores for whole and tumor core, they do so at the expense of either complexity (McKinley [43]) or a large number of parameters (about 52M in Isensee *et al.* [10]). In contrast to our proposed study, Zhou *et al.* [27] found that using a range of attention mechanisms, network ensembling architectures, and a post-processing step has resulted in improved scores of the whole tumor and tumor core. The combined approaches solved the complex difficulties of cascaded UNet. However, the ensembling strategy has demanded massive computing storage resources for storing the 3D medical datasets. As a result, a single best architecture, such as ours, would be desired in such high-demand computing situations. Kermi *et al.* [44] presented a residual-based 2D UNet variant. The current network has been trained with a weight loss function to balance majority and minority classes.

TABLE 5. The comparison of different approaches’ performances on the BraTS 2020 validation dataset. The metrics’ scores are reported as a mean. The best value of the scores is highlighted in bold.

Methods	Dice Scores			Hausdorff95(mm)		
	WT	TC	ET	WT	TC	ET
Jia <i>et al.</i> [Cascaded Ensemble] [47]	91.022	85.701	77.338	4.302	4.933	29.712
Henry <i>et al.</i> [Ensemble] [48]	91.123	84.921	72.738	4.301	5.692	20.558
Vu <i>et al.</i> [Ensemble (7 models)] [40]	90.550	82.670	77.170	4.990	8.630	27.040
Fidon <i>et al.</i> [Ensemble] [41]	91.000	84.400	77.600	4.400	5.800	26.800
Ghaffari <i>et al.</i> [Ensemble] [13]	90.000	82.000	78.000	—	—	—
(Proposed)	90.639	83.624	78.288	4.164	9.809	32.200

Compared to 2D technique (10 M) of Kermi *et al.* [44], our proposed approach with the non-weight loss functions learns fewer parameters (3.7 M) for the best scores. Albiol *et al.* [Ensemble] [45] describes a method for ensembling different 3D CNN architectures. The 3D convolutions replace the 2D convolutions of several architectures such as VGG (Albiol *et al.* [VGG-like] [45]), inception versions 2 and 3 (Albiol *et al.* [Inception2, Inception3] [45]), densely connected networks (Albiol *et al.* [Densely Connected] [45]). However, in terms of mean dice scores for all tumors, our proposed method outperformed all single 3D and ensemble techniques. Furthermore, Feng *et al.* [Ensemble] [46] describes an ensemble technique. The average function is used for six different 3D UNet versions that have been trained with both weight and non-weight loss functions. All of the aforementioned techniques made use of more initial training channels. On the other hand, our proposed model is the lightest, most in-depth, and can achieve competitive scores. Our method has a drawback for metric Hausdroff95 distances since the average scores of the three tumors are not decreased. In addition, Figure 11 depicts the proposed architecture’s segmentation results.

3) MICCAI BRATS 2020 CHALLENGE

In terms of metrics, like DSC, and Hausdroff95 distances, Table 5 presents a comparison between the proposed model and state-of-the-art approaches. Only the mean DSC scores are taken into account for the best techniques. Table 5 shows the best mean dice score of the whole tumor using ensemble technique of Henry *et al.* [Ensemble] [48]. Meanwhile, after executing an ensemble operation on five models, a cascaded technique [47] delivers the best mean value of the tumor core. At the same time, our proposed single model has the highest value of the enhancing tumor.

Jia *et al.* [47] achieved the best mean value of the tumor core by combining diverse strategies in a 3D UNet architecture, including residual networks for better feature utilization, numerous local and global context fusion blocks, and an attention block. Jia *et al.* [Cascaded Ensemble] [47] provided a multi-step approach. The output of the first step is merged with input modalities as the input for the second step in the multi-step solution. The existing two-step method requires a

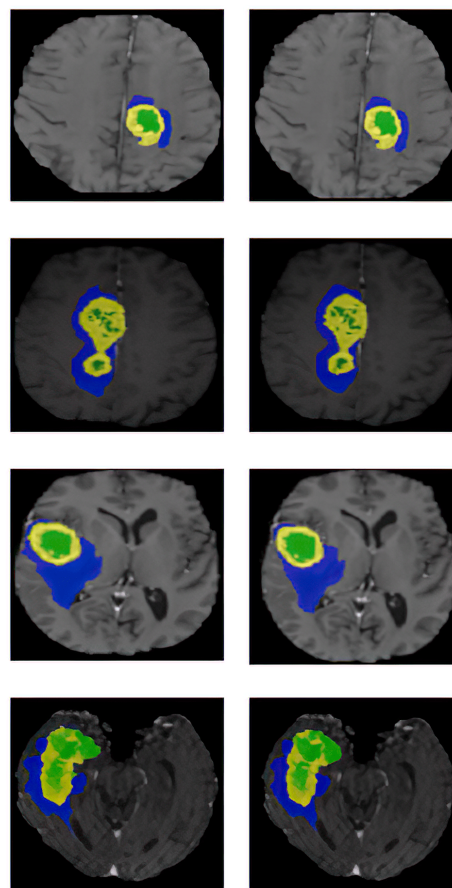


FIGURE 11. Instance segmentation results on the BraTS 2018 training dataset. From left to right: truth-label and proposed model predictions overlaid on T1ce modality, whole tumor (blue), tumor core (green), and enhancing tumor (yellow).

large amount of memory. Additionally, the two-step solution has learned approximately 26 M parameters. Furthermore, an ensemble process is applied to the ten two-step approaches for the best score of the tumor core. In contrast, our single proposed approach learns only 3.7 M parameters for the best score of the enhancing tumor. For the best score of the whole tumor, an ensemble technique (Henry *et al.* [Ensemble] [48]) is used on the five models of the original 3D UNet variant.

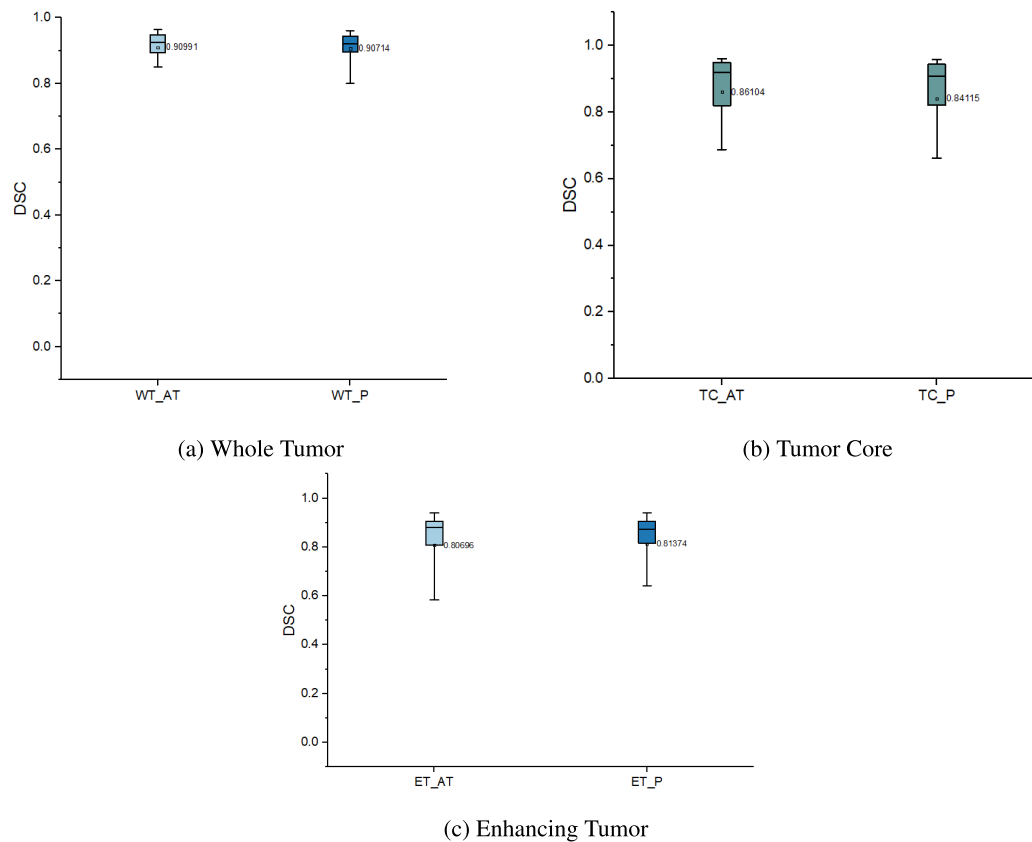


FIGURE 12. DSC comparison. From a to c, individual tumors of MH UNet with attention strategies (denoted by WT_AT, TC_AT, and ET_AT) and MH UNet (denoted by WT_P, TC_P, and ET_P) are compared. The numbers at the boxes are mean average DSC values.

This variation, however, is trained with 48 initial channels. In contrast, our proposed merely trains with 16 initial channels for the best score of the enhancing tumor. Vu *et al.* [Ensemble (7 models)] [40] describes a cascaded technique that comprises many UNets. The current cascaded technique is a 3D UNet variant in itself. This variant, like regular 3D UNet, has an encoder and decoder. The decoder, however, comprises three distinct 3D UNets. Each 3D UNet is used to segment the individual tumors. Finally, the final scores are reported by combining the segmented tumors of three 3D UNets. Furthermore, the predicted MRIs of seven such cascaded approaches are subjected to an ensemble operation. As a result, the cascaded approaches increase the complexity of solving the segmentation problem. Fidon *et al.* [Ensemble] [41] describes an ensemble approach for 3D UNets. The original 3D UNet is trained with different loss functions, including the sophisticated weight hyper-parameter tuning loss function and optimizers. Finally, the predictions of several 3D UNets are combined. Ghaffari *et al.* [Ensemble] [13] also mentions an ensembling approach. The original 3D UNet is modified by incorporating residual learning, dense networks, and deep supervision. The current method employs a single 3D UNet and a cascade of 3D UNets to segment the tumors. Finally, the predictions of single 3D UNet and

cascaded 3D UNets are combined for the final scores. Compared to ensembles of various models such as Vu *et al.* [Ensemble (7 models)] [40] and Ghaffari *et al.* [Ensemble] [13], our proposed technique achieves the best mean scores for all tumors. When compared to all other techniques, our proposed architecture, in particular, can secure the best mean score for at least one tumor. Furthermore, the Hausdorff95 distances for the whole tumor have the best mean score for our proposed architecture. One disadvantage of our proposed work is that it has a lower score for tumor core. However, the ensemble approaches outlined in contemporary techniques may improve the lower score of the tumor core.

We will, however, use the single proposed architecture to improve the lower score of tumor core by including numerous attention strategies [49]. As shown in Figure 12b, the average DSC score of the tumor core increases by 19.89% using the scales, channels, and positions attention strategies in MH UNet. Meanwhile, numerous attention strategies enable MH UNet to learn only essential foreground features for improved whole tumor score (see Figure 12a). However, we observe a 6.7% lower score in the enhancing tumor than the proposed MH UNet (see Figure 12c). Therefore, novel attention strategies should be investigated appropriately to improve the mean DSC scores of all brain tumors.

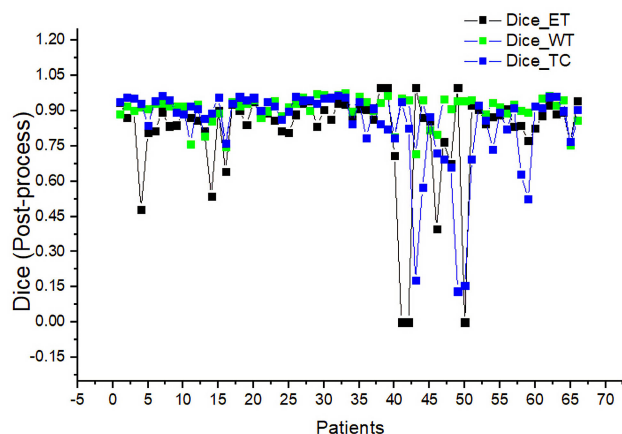


FIGURE 13. Graphical representation of DSC. In the graph, X-axis represents the 66 patients of the BraTS 2018 validation dataset, while Y-axis shows DSC (post-processed) values.

To eliminate false-positive voxels in all BraTS competitions, we implement a post-processing step to the training and validation predictions. For the BraTS 2018 validation dataset, we plot the mean dice scores of all tumors. The graph of the post-process DSC metric is shown in Figure 13. In *LGG* cases, we still witness a zero dice score for the enhancing tumor (denotes by black) (see patients between 40 to 55 on the X-axis). Better post-processing techniques [27], [28], [38], [50] are therefore necessary to reduce these false-positive voxels.

As shown in Figure 11, our proposed MH UNet segments the whole tumor area accurately. MH UNet also finds the enhancing tumor and tumor core, which are small and difficult to detect. Due to the small size of the enhancing tumor and tumor core and the varied locations of tumors, MH UNet still generates some false-positives (white arrow) and false-negatives (black-arrows), as shown in Figure 14c. Meanwhile, by incorporating some strategies in MH UNet, the problem of false-positives and false-negatives can minimize. To minimize over-fitting during training, we use augmentation techniques, including random rotation and flipping. However, augmentation techniques can also help improve segmentation accuracy during testing [10], as shown in Figure 14d. Despite the superiority of test-time augmentation in removing the false-positives and false-negatives, still, we observe false-positives in the top image of Figure 14d. These false-positives are further minimized when we train the MH UNet with the higher image resolutions and more initial channels (24), as shown in Figure 14e. However, false-positives exist in the top image of Figure 14e. On ensemble the test-time augmented predictions of the proposed model, the results perfectly match the ground-truths, as shown in Figure 14f. However, the ensembling of several models raises storage issues and increases the complexity of solving the segmentation tasks. In the future, we will try to remove the false-positives and false-negatives from all predicted cases without ensemble approaches.

V. ISLES 2018 CHALLENGE

We also use our MH UNet to segment stroke lesions using challenging ISLES 2018 training and test sets.

A. DATASET

Normal functioning of the brain relies on the sufficient supply of blood oxygen through arteries and veins. Often blood flow is obstructed, causing tissue death. The dead tissue with an area is known as a stroke lesion. Stroke is a life-threatening condition, often called cerebrovascular disease. Segmented stroke lesion diagnosis may assist with evaluation and treatment planning. Thus, automated segmentation of stroke lesions is an optimal practice for accurate details.

Although neurologists use the *computed tomography* (CT) technique to obtain precise brain stroke, CT scans of patients with indistinct information are not suitable for automated methods. Our work provides an automatic, lightest method for accurately segmenting stroke lesions using CT perfusion images. *Ischemic Stroke Lesion Segmentation* (ISLES) 2018 challenge has training and testing datasets for competition. We used 94 cases of training dataset and 62 cases of testing dataset [51], [52] for our proposed work. Each case includes many CT perfusion modalities such as *cerebral blood volume* (CBV), *cerebral blood flow* (CBF), *residue peak time* (Tmax), *mean transit time* (MTT) and CT. The input is created from these 5 modalities. Furthermore, the truth-labels (generated using MRI *Diffusion-Weighted Imaging* (DWI)) are given with the training dataset, while there is no truth-label for test cases. The predictions of the training and the testing sets are submitted for the final evaluation.

B. DATA PRE-PROCESSING

A bias correction step is performed on each case of the training and testing dataset. We also normalized each case of the training dataset. We extracted the patches of size $128 \times 128 \times 32$ from modalities.

C. IMPLEMENTATION DETAILS

The training set (total cases 94) is divided into five parts. Thus the MH UNet is trained and tested five times. Every time, there are 76 cases for training and the remaining patients for the validation. During the training, we use optimizer Adam with a batch size of 4. The MH UNet is trained with an initial learning rate of 5×10^{-4} , which drops by 50% if validation loss is not improved within 30 epochs. The MH UNet is trained for 300 epochs. In addition, augmentation techniques such as random rotation and flipping are used to avoid over-fitting during training. During the networks' training, we use the loss function of Equation (15). However, the combined binary loss function for the ISLES 2018 training set replaces this multi-label loss function.

D. EVALUATION

We evaluate the predictions of the ISLES dataset. *Dice Similarity Coefficient* (DSC) or Dice, Accuracy, Recall, Hausdorff Distance, Average Distance, and Absolute Volume Difference

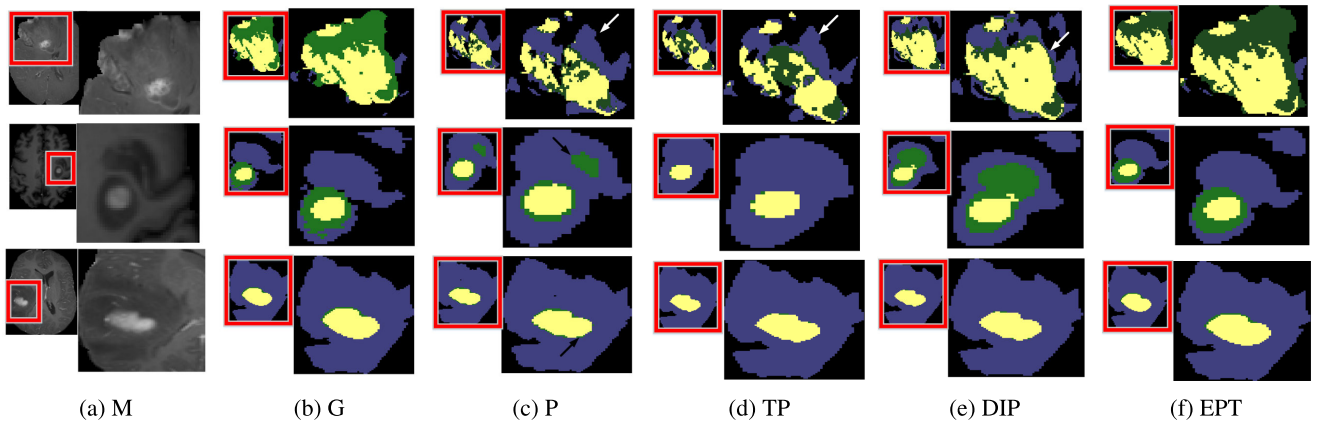


FIGURE 14. Illustration of BraTS 2018 dataset's segmentation results. Each column from (a) to (f) shows images and their enlarged versions. (a) and (b) respectively show the T1ce modalities (M) and ground-truths (G). (c) shows the worst results of the proposed model (P). (d) - (f) show possible improvements of the proposed model. The false-positives (denoted by a white arrow) and false-negatives (denoted by the black arrows) of P minimize when a test-time augmentation strategy applies to the predictions of the proposed model (TP). The predictions are further improved when P is trained with larger input resolutions and more initial channels (DIP). The larger input resolutions are possible by reducing the depth (D). As a result, the number of initial channels (I) increases to train the modified P. On ensemble (E), the test-time augmented predictions (T) from the five variations of the proposed model (P), the outcomes (EPT) exactly match the ground-truths (G). Each colour represents a different tumor: green for the tumor core, blue for the whole tumor, and yellow for enhancing the tumor.

TABLE 6. The comparison of different approaches' performances on the ISLES 2018 training dataset. The scores are provided as mean. The average scores in bold are written from the organizer's verification.

	Dice	Hausdorff Distance	Average Distance	Precision	Recall	AVD
Bertels <i>et al.</i> [21]	0.75	24.17	1.97	0.70	0.84	7.34
Islam <i>et al.</i> [22]	0.74	22.40	1.81	0.67	0.84	8.16
Tureckova <i>et al.</i> [24]	0.71	20.13	1.69	0.64	0.82	9.92
Proposed	0.82	17.09	0.68	0.77	0.90	5.61

or AVD are used in the evaluation of the predicted labels. Each metric is defined properly in Taha and Hanbury [53].

E. COMPARISON WITH THE BASELINE APPROACHES

In this sub-section, our proposed work is compared with the following baseline approaches: Bertels *et al.* [21], Islam *et al.* [22], and Tureckova and Rodríguez-Sánchez [24]. These approaches are discussed in the subsequent sub-sections. We use ISLES 2018 training dataset (94 cases) for all approaches. For each approach, 76 ISLES patients are available for training and 18 for validation. Each approach is then used to evaluate all ISLES training cases (94) once it has been trained. Table 6 shows the mean scores of all metrics for each approach. The depth and multi-scale aspects of the proposed work allow it to acquire the best mean scores of all metrics.

The proposed work has precisely segmented the stroke lesion when the predictions of several approaches are visualized, as seen in the first row of Figure 15. Meanwhile, Tureckova and Rodríguez-Sánchez [24] generate a mis-segmentation outcome (a green arrow), while Bertels *et al.* [21] and Islam *et al.* [22] have incorrectly predicted stroke lesions. A drawback of our proposed work is a mis-segmentation result, as seen in the last row of Figure 15. The mis-segmentation outcomes highlight the difficulty that different architectures face when attempting to obtain exact

segmentation results. Nevertheless, based on the highest mean scores of several metrics, our proposed work is better than the Bertels *et al.* [21], Islam *et al.* [22], and Tureckova and Rodríguez-Sánchez [24].

F. RESULTS OF MICCAI ISLES 2018 CHALLENGE

1) RESULTS OF MICCAI ISLES 2018 TRAINING DATASET

Table 7 shows the mean scores for each metric from various users, including our MH UNet. After comparing MH UNet with other ISLES 2018 training users (see section Leaderboard: Training ¹), the mean scores for each metric are reported. In Table 7, our proposed work outperformed state-of-the-art approaches in terms of DSC, Average Distance, and Recall metrics. Meanwhile, the mean scores of Hausdorff Distance, Precision, and AVD in our proposed work are lower. Some post-processing techniques, such as uncertainty filtering [54], may boost the mean scores of these metrics. Figure 16 shows the segmentation results of the proposed model.

Nonetheless, the DSC metric, a decisive metric for the best approaches, reports the best mean scores in our proposed work. We also give the mean DSC scores of various approaches [20]–[24] in addition to the training users of Table 7. A GAN technique based on the 2D variant of UNet

¹(<https://www.smir.ch/ISLES/Start2018>)

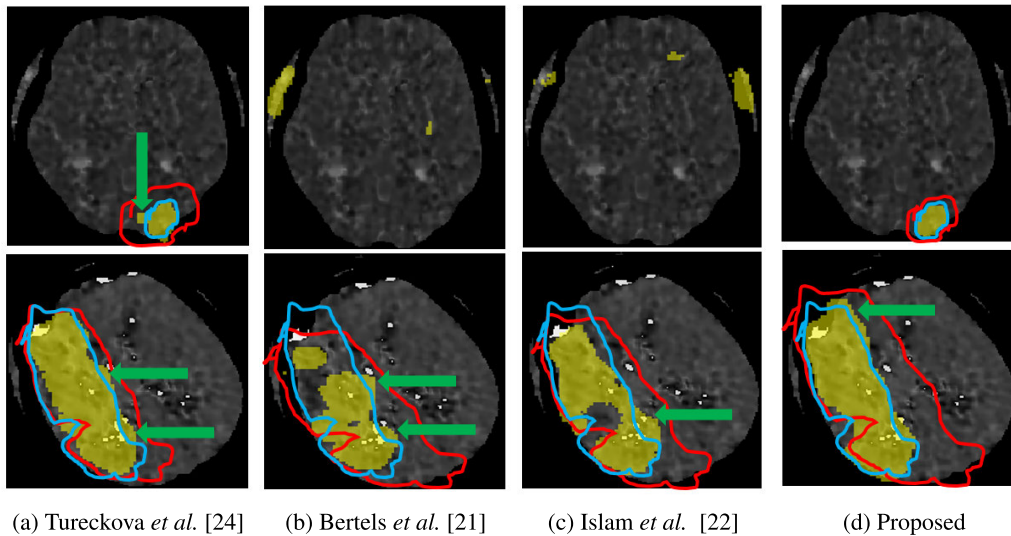


FIGURE 15. On the ISLES 2018 dataset, a visual comparison of the different approaches for ischemic stroke lesion segmentation. Truth values and predictions respectively show by blue and red curves. Lesion shows in yellow. Green arrows denote mis-segmentations.

TABLE 7. The comparison of different training users’ performances on the ISLES 2018 training dataset. The scores are presented as mean (standard deviation). The best mean scores are highlighted in bold and are written from the organizer’s verification.

	Dice	Hausdorff Distance	Average Distance	Precision	Recall	AVD
lilis2	0.80 (0.17)	13.68 (14.36)	0.99 (2.67)	0.84 (0.15)	0.78 (0.20)	2.57 (4.70)
zhens3	0.80 (0.19)	16.82 (18.52)	1.34 (3.27)	0.79 (0.19)	0.83 (0.19)	2.94 (4.28)
zhengd1	0.78 (0.16)	1063847.90 (10259201.06)	1063830.69 (10259201.84)	0.79 (0.16)	0.78 (0.17)	2.75 (3.21)
zhany21	0.72 (0.24)	2127679.45 (14430486.40)	2127664.57 (14430488.59)	0.73 (0.22)	0.74 (0.27)	3.76 (3.82)
meimc1	0.70 (0.17)	19.22 (19.24)	2.00 (4.55)	0.64 (0.19)	0.84 (0.20)	11.40 (17.68)
zhans10	0.70 (0.18)	1063849.19 (10259200.92)	1063831.29 (10259202.78)	0.74 (0.18)	0.71 (0.22)	7.72 (12.79)
pinhg2	0.64 (0.21)	26.56 (24.52)	3.21 (7.83)	0.61 (0.24)	0.75 (0.19)	9.46 (13.00)
ahmap3 (Proposed)	0.82 (0.08)	17.09 (18.38)	0.68 (0.96)	0.77 (0.11)	0.90 (0.07)	5.61 (7.40)

was developed by Liu [20]. The mean dice score for this GAN technique was 61. For the mean dice score of 42, Islam *et al.* [22] proposes an alternative GAN technique based on multiple 2D UNets. Bertels *et al.* [21] proposed a shallow encoder-decoder design that was similar to original UNet [7] for the mean dice score of 49 and included a weight loss function. Dolz *et al.* [19] demonstrated a 2D densely equipped UNet architecture that took inputs from several modalities. Using numerous fusion procedures, the features of different modalities can be represented in meaningful ways. Finally, the densely equipped UNet has achieved a 64 average dice score. Most 2D variants of UNet are employed in the following methodologies. However, for stroke lesion segmentation, some 3D variants of UNet have been proposed. Pinheiro *et al.* [23] proposed a 3D variation of UNet [7]. The network, though, is shallow. The residual connections added to the encoder of the 3D UNet [24] improve this shallow network. As an element-wise addition operation is conducted to the previous and next layer features, residual connections can (1) increase the depth to existing encoder-decoder architectures and (2) improve representations. The residual-based variations of UNet, on the other hand, learn more parameters than the dense-based UNet.

Existing UNet variants have limited depth and multi-scale capabilities. This difficulty can be mitigated by adhering to our architecture’s design. Furthermore, contemporary UNet techniques solve the high-class imbalance issue by using appropriate weight factors that are manually selected for distinct weight loss functions. On the other hand, the time spent manually searching for the most critical weight factors is crucial. Non-weight loss functions, such as those we propose in our paper, can further minimise the problem.

2) RESULTS OF MICCAI ISLES 2018 TESTING DATASET

All ISLES 2018 testing (62) cases are evaluated using the MH UNet. All predicted labels are submitted online² for final assessment. Table 8 shows a comparison of our work to state-of-the-art techniques [21], [22], [24] in terms of performance. Our work resulted in the highest average DSC score. At the same time, a 2D variant of UNet [22] reports the same mean DSC score as ours. However, Islam *et al.* [22] adopted a GAN technique based on several UNets. That is, the given GAN technique is utilized to segment the ISLES dataset in more than one step. Meanwhile, the stroke lesions are segmented

²(<https://www.smir.ch/ISLES/>)

TABLE 8. The comparison of different approaches' performances on the ISLES 2018 testing dataset. The scores are presented as mean (standard deviation). The best mean value of each metric is highlighted in bold.

Methods	Dice	Hausdorff Distance	Average Distance	Precision	Recall	AVD	Parameters
Bertels <i>et al.</i> [21]	0.38 (0.30)	– (–)	– (–)	0.47 (0.35)	0.44 (0.34)	17.2 (16.9)	–
Islam <i>et al.</i> [22]	0.39 (–)	17.74 (–)	17.74 (–)	0.55 (–)	0.36 (–)	10.90 (–)	–
Tureckova <i>et al.</i> [24]	0.37 (–)	19.35 (–)	19.35 (–)	0.44 (–)	0.44 (–)	24.95 (–)	8.20 M
ahmap3 (Proposed)	0.39 (0.29)	16.13 (36.77)	16.13 (36.77)	0.52 (0.38)	0.37 (0.29)	13.48 (17.38)	3.7 M

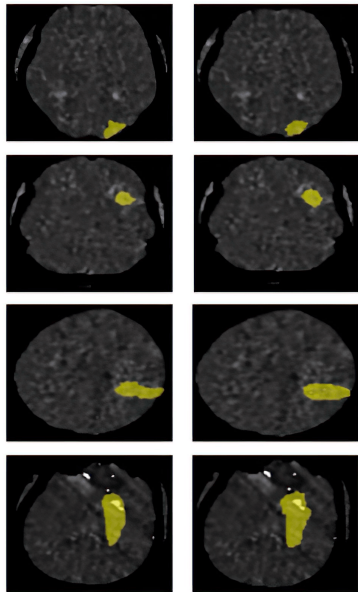


FIGURE 16. Instance segmentation results on the ISLES 2018 training dataset. From left to right: truth-label, proposed model predictions overlaid on MTT modality. Lesion is shown in yellow.

in a single step using our proposed method. Simultaneously, a 3D variant of UNet [24] obtained the best mean recall score, albeit with 2.2 times the number of parameters as our work. Furthermore, we are able to secure the best position in hausdorff and average distances. The incorrect number of slices for the third dimension is a key flaw in the ISLES training dataset. In the z dimension, case number 8 in the ISLES training dataset, for example, comprises just eight slices. As a result, creating any variant of UNet that can achieve the best mean scores for all metrics on the testing dataset remains difficult. We will be concentrating on this rugged design in the future. In addition, with our proposed work, we will try test-time augmentations to eliminate false-positive voxels. In addition, we will test the majority ensembling technique on numerous different UNet architectural modifications to improve test scores.

VI. CONCLUSION

This paper proposes a variant form of a 3D UNet for medical image segmentation. To reduce the training parameters and efficient gradient flow, we implemented densely connected blocks in the proposed MH UNet. Simultaneously, dense connections used the minimal growth-rate value to

remove unnecessary convolution layer features. As a result, we addressed the issue of huge learnable parameters. The MH UNet also used residual-inception blocks to learn multi-scale contexts. In encoder-decoder, we proposed two variations of residual-inception blocks. Furthermore, we proposed a hierarchical block that incorporates the various parallel dilated convolution layers to expand the size of the limited receptive field in the feature maps of the dense blocks at the encoder. Additionally, we employed a deep supervision approach for faster convergence and superior segmentation accuracy. Simultaneously, the deep supervision approach enhances segmentation accuracy by combining various depths' segmentation maps. The MICCAI BraTS and ISLES datasets are used to check the performances of the MH UNet. The proposed MH UNet achieved considerable segmentation scores on the BraTS dataset. Meanwhile, our MH UNet achieved competitive segmentation scores on the ISLES 2018 testing dataset. In the future, we will apply effective post-processing algorithms to improve the performance of medical datasets. In conclusion, we believe that our proposed approach would achieve state-of-the-art performance on other challenging medical datasets.

VII. SUPPLEMENTARY MATERIALS

A. MH UNet's LAYERS' VISUALIZATION AND MH UNet's DETAILS

We visualize some layers of our proposed work in this part. Individual tumors for a single layer are visualized in each row of Tables 9, 10, 11. In each row of the table, from left to right: truth-labels and single layer's predictions (of BraTS 2018 dataset) overlaid on T1ce modality, enhancing tumor (yellow), tumor core (green), and whole tumor (blue).

The same BraTS 2018 patient is visualized for the different layers in Table 9. The output feature maps of ten 3D convolution layers are: 16×128^3 , 16×128^3 , 2×128^3 , 2×128^3 , 2×128^3 , 16×128^3 , 16×128^3 , 16×128^3 , 2×128^3 , and 2×128^3 . Except for rows 4 and 5, all layers indicate 0 mean dice scores for the whole tumor. The tumor core has a similar scenario but for the different layers. Meanwhile, in every layer except row 1, the small size of an enhancing tumor is quite accurately predicted. At least one tumor can be predicted by row numbers 4 to 6. The proposed residual-inception block, in which dense networks further enhance output features of several dilated layers, supports this.

Table 10 depicts different BraTS 2018 patients in comparison to Table 9. Each row represents a separate layer.


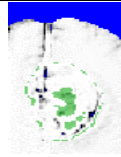
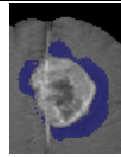
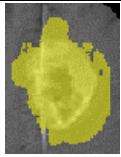

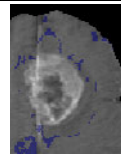
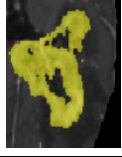
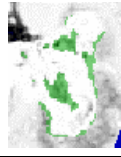
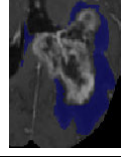
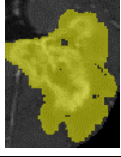
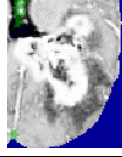
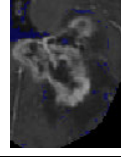
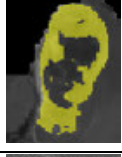

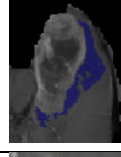


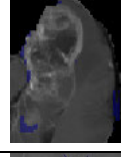
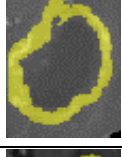

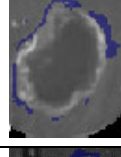

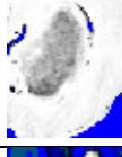
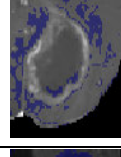
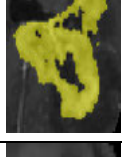
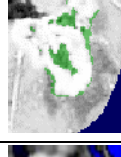
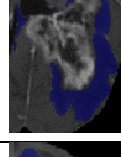


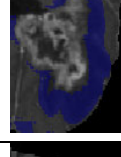


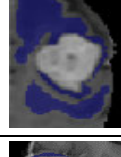


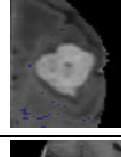


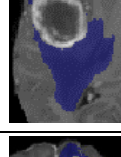
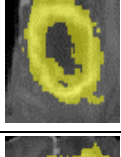

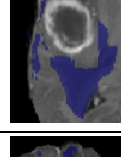
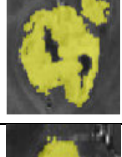
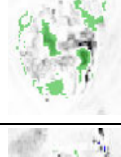
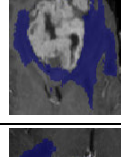
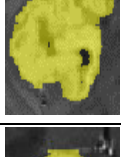
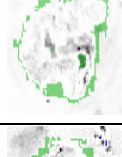
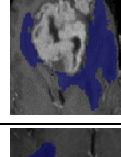
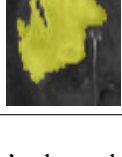
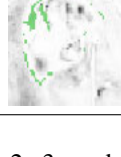
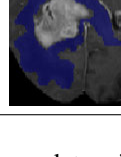

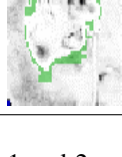
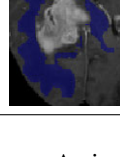
TABLE 9. On the BraTS 2018 dataset, a visualization of the various layers for brain tumor segmentation. In each row, from left to right: truth-labels and single layer’s predictions are overlaid on T1ce modality, enhancing tumor (yellow), tumor core (green), and whole tumor (blue).

Ground Truth			Prediction		
ET	TC	WT	ET	TC	WT

TABLE 10. On the BraTS 2018 dataset, a visualization of the various layers for brain tumor segmentation. In each row, from left to right: truth-labels and single layer's predictions are overlaid on T1ce modality, enhancing tumor (yellow), tumor core (green), and whole tumor (blue).

Ground Truth			Prediction		
ET	TC	WT	ET	TC	WT

TABLE 11. On the BraTS 2018 dataset, a visualization of the various layers for brain tumor segmentation. In each row, from left to right: truth-labels and single layer’s predictions are overlaid on T1ce modality, enhancing tumor (yellow), tumor core (green), and whole tumor (blue).

Ground Truth			Prediction		
ET	TC	WT	ET	TC	WT
					
					
					
					
					
					
					
					
					

The encoder’s dense blocks 2, 3, and 4 are used to pick these layers. In the maximum number of rows, the mean dice score of the whole tumor for the BraTS 2018 patients is zero. Similarly, row numbers 6 and 10 accurately predict the whole tumor. In the meantime, the enhancing tumor with zero dice

scores for rows 1 and 2 can be seen. A similar scenario can be seen in the tumor core, albeit on different rows. Table 10 shows how the location, shape, and size of the various tumors differ. Nonetheless, each tumor type can be predicted by the layers of our proposed model.

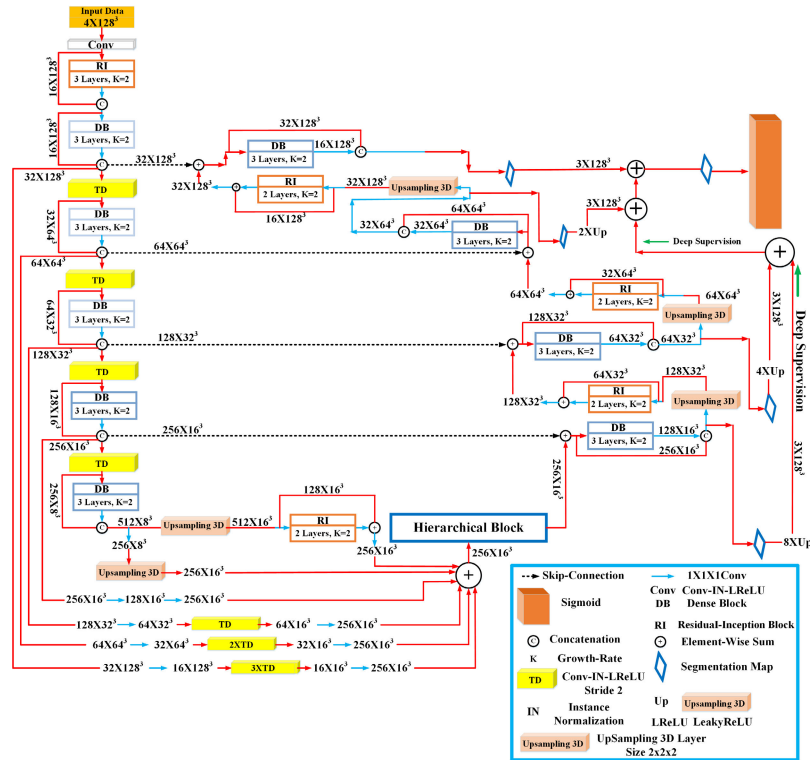


FIGURE 17. MH UNet in details.

In Table 11, we select layers of the hierarchical and decoder blocks for the visualization. In comparison to Table 9 and Table 10, we have chosen a variety of BraTS 2018 patients. The purpose of choosing different patients is to assess the potential of the proposed architecture when the tumor’s location, shape, and size varies. Table 11 shows accurate predictions of the whole tumor. However, in some instances, a zero dice score for the whole tumor still exists. In the case of the tumor core, the scenario is similar. Meanwhile, larger receptive field sizes of the deeper layers in terms of the accurate predictions of the enhancing tumor for most rows have been observed.

We have depicted some layers of our proposed architecture in each of the tables above. At least one predicted tumor should be assumed for the best results of each layer. The worst-case scenario, in which each layer has only one predicted tumor, should be assumed. However, the potential of each layer for each BraTS 2018 patient cannot be exhibited because of space constraints. Nonetheless, for all BraTS datasets, our proposed architecture has the best mean dice scores of the enhancing tumor. Meanwhile, our proposed architecture has nearly similar dice scores for the tumor core and the whole tumor, like state-of-the-art techniques.

ACKNOWLEDGMENT

This work was supported by the Taif University Researchers Supporting Project, Taif University, Taif, Saudi Arabia, under Grant TURSP-2020/36.

REFERENCES

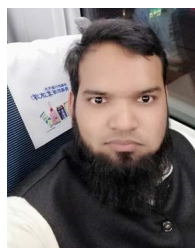
- [1] S. Qamar, H. Jin, R. Zheng, and P. Ahmad, “3D hyper-dense connected convolutional neural network for brain tumor segmentation,” in *Proc. 14th Int. Conf. Semantics, Knowl. Grids (SKG)*, Sep. 2018, pp. 123–130.
- [2] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” 2016, *arXiv:1603.05959*. [Online]. Available: <http://arxiv.org/abs/1603.05959>
- [3] S. Qamar, H. Jin, R. Zheng, and P. Ahmad, “Multi stream 3D hyper-densely connected network for multi modality isointense infant brain MRI segmentation,” *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 25807–25828, Sep. 2019.
- [4] J. N. Sua, S. Y. Lim, M. H. Yulius, X. Su, E. K. Y. Yapp, N. Q. K. Le, H.-Y. Yeh, and M. C. H. Chua, “Incorporating convolutional neural networks and sequence graph transform for identifying multilabel protein lysine PTM sites,” *Chemometric Intell. Lab. Syst.*, vol. 206, Nov. 2020, Art. no. 104171. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169743920303774>
- [5] N. Q. K. Le, Q.-T. Ho, E. K. Y. Yapp, Y.-Y. Ou, and H.-Y. Yeh, “DeepETC: A deep convolutional neural network architecture for investigating and classifying electron transport chain’s complexes,” *Neurocomputing*, vol. 375, pp. 71–79, Jan. 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231219313384>
- [6] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, “Brain tumor segmentation with deep neural networks,” 2015, *arXiv:1505.03540*. [Online]. Available: <http://arxiv.org/abs/1505.03540>
- [7] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” 2015, *arXiv:1505.04597*. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [8] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, “Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge,” 2018, *arXiv:1802.10508*. [Online]. Available: <http://arxiv.org/abs/1802.10508>

- [9] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Med. Image Anal.*, vol. 51, pp. 21–45, Jan. 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S136184151830848X>
- [10] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," 2018, *arXiv:1809.10483*. [Online]. Available: <http://arxiv.org/abs/1809.10483>
- [11] W. Wang, C. Chen, M. Ding, J. Li, H. Yu, and S. Zha, "Trans-BTS: Multimodal brain tumor segmentation using transformer," 2021, *arXiv:2103.04430*. [Online]. Available: <http://arxiv.org/abs/2103.04430>
- [12] G. Wang, T. Song, Q. Dong, M. Cui, N. Huang, and S. Zhang, "Automatic ischemic stroke lesion segmentation from computed tomography perfusion images by image synthesis and attention-based deep neural networks," 2020, *arXiv:2007.03294*. [Online]. Available: <http://arxiv.org/abs/2007.03294>
- [13] M. Ghaffari, A. Sowmya, and R. Oliver, "Brain tumour segmentation using cascaded 3D densely-connected U-Net," 2020, *arXiv:2009.07563*. [Online]. Available: <http://arxiv.org/abs/2009.07563>
- [14] U. Baid, N. A. Shah, and S. Talbar, "Brain tumor segmentation with cascaded deep convolutional neural network," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. Cham, Switzerland: Springer, 2020, pp. 90–98.
- [15] M. H. Vu, T. Nyholm, and T. Löfstedt, "TuNet: End-to-end hierarchical brain tumor segmentation using cascaded networks," 2019, *arXiv:1910.05338*. [Online]. Available: <http://arxiv.org/abs/1910.05338>
- [16] S. Kim, M. Luna, P. Chikontwe, and S. H. Park, "Two-step U-Nets for brain tumor segmentation and random forest with radiomics for survival time prediction," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. Cham, Switzerland: Springer, 2020, pp. 200–209.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
- [19] J. Dolz, I. B. Ayed, and C. Desrosiers, "Dense multi-path U-Net for ischemic stroke lesion segmentation in multiple image modalities," 2018, *arXiv:1810.07003*. [Online]. Available: <http://arxiv.org/abs/1810.07003>
- [20] P. Liu, "Stroke lesion segmentation with 2D novel CNN pipeline and novel loss function," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijff, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 253–262.
- [21] J. Bertels, D. Robben, D. Vandermeulen, and P. Suetens, "Contra-lateral information CNN for core lesion segmentation based on native CTP in acute stroke," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijff, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 263–270.
- [22] M. Islam, N. R. Vaidyanathan, V. J. M. Jose, and H. Ren, "Ischemic stroke lesion segmentation using adversarial learning," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijff, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 292–300.
- [23] G. R. Pinheiro, R. Voltoline, M. Bento, and L. Rittner, "V-Net and U-Net for ischemic stroke lesion segmentation in a small dataset of perfusion data," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijff, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 301–309.
- [24] A. Tureckova and A. J. Rodríguez-Sánchez, "ISLES challenge: U-shaped convolution neural network with dilated convolution for 3D stroke lesion segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijff, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 319–327.
- [25] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," 2018, *arXiv:1810.11654*. [Online]. Available: <http://arxiv.org/abs/1810.11654>
- [26] P. Ahmad, H. Jin, S. Qamar, R. Zheng, and A. Saeed, "RD2A: Densely connected residual networks using ASPP for brain tumor segmentation," *Multimedia Tools Appl.*, vol. 80, pp. 27069–27094, May 2021.
- [27] C. Zhou, C. Ding, X. Wang, Z. Lu, and D. Tao, "One-pass multi-task networks with cross-task guided attention for brain tumor segmentation," 2019, *arXiv:1906.01796*. [Online]. Available: <http://arxiv.org/abs/1906.01796>
- [28] S. Wang, C. Dai, Y. Mo, E. Angelini, Y. Guo, and W. Bai, "Automatic brain tumour segmentation and biophysics-guided survival prediction," 2019, *arXiv:1911.08483*. [Online]. Available: <http://arxiv.org/abs/1911.08483>
- [29] M. Islam, V. S. Vibashan, V. J. M. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain tumor segmentation and survival prediction using 3D attention UNet," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. Cham, Switzerland: Springer, 2020, pp. 262–272.
- [30] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. Yuille, and L. Fei-Fei, "Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation," 2019, *arXiv:1901.02985*. [Online]. Available: <http://arxiv.org/abs/1901.02985>
- [31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014, *arXiv:1409.4842*. [Online]. Available: <http://arxiv.org/abs/1409.4842>
- [32] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: <http://arxiv.org/abs/1602.07261>
- [33] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," 2017, *arXiv:1702.08502*. [Online]. Available: <http://arxiv.org/abs/1702.08502>
- [34] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," 2016, *arXiv:1607.00582*. [Online]. Available: <http://arxiv.org/abs/1607.00582>
- [35] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, Art. no. 170117, Sep. 2017, doi: [10.0.4.14/sdata.2017.117](https://doi.org/10.0.4.14/sdata.2017.117).
- [36] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, and L. Lanczi, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.
- [37] Y.-X. Zhao, Y.-M. Zhang, and C.-L. Liu, "Bag of tricks for 3D MRI brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. Cham, Switzerland: Springer, 2020, pp. 210–220.
- [38] R. McKinley, M. Rebsamen, R. Meier, and R. Wiest, "Triplanar ensemble of 3D-to-2D CNNs with label-uncertainty for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. Cham, Switzerland: Springer, 2020, pp. 379–387.
- [39] D. Lachinov, E. Shipunova, and V. Turlapov, "Knowledge distillation for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, vol. 11993, A. Crimi and S. Bakas, Eds. Cham, Switzerland: Springer, 2019, pp. 324–332, doi: [10.1007/978-3-030-46643-5_32](https://doi.org/10.1007/978-3-030-46643-5_32).
- [40] M. H. Vu, T. Nyholm, and T. Löfstedt, "Multi-decoder networks with multi-denoising inputs for tumor segmentation," 2020, *arXiv:2012.03684*. [Online]. Available: <http://arxiv.org/abs/2012.03684>
- [41] L. Fidon, S. Ourselin, and T. Vercauteren, "Generalized wasserstein dice score, distributionally robust deep learning, and ranger for brain tumor segmentation: BraTS 2020 challenge," 2020, *arXiv:2011.01614*. [Online]. Available: <http://arxiv.org/abs/2011.01614>
- [42] M. Chen, Y. Wu, and J. Wu, "Aggregating multi-scale prediction based on 3D U-Net in brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. Cham, Switzerland: Springer, 2020, pp. 142–152.
- [43] R. McKinley, R. Meier, and R. Wiest, "Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijff, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 456–465.
- [44] A. Kermi, I. Mahmoudi, and M. T. Khadir, "Deep convolutional neural networks using U-Net for automatic brain tumor segmentation in multimodal MRI volumes," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijff, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 37–48.
- [45] A. Albiol, A. Albiol, and F. Albiol, "Extending 2D deep learning architectures to 3D image segmentation problems," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, vol. 11384, A. Crimi, S. Bakas, H. J. Kuijff, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2018, pp. 73–82, doi: [10.1007/978-3-030-11726-9_7](https://doi.org/10.1007/978-3-030-11726-9_7).

- [46] X. Feng, N. Tustison, and C. Meyer, "Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using radiomic features," 2018, *arXiv:1812.01049*. [Online]. Available: <http://arxiv.org/abs/1812.01049>
- [47] H. Jia, W. Cai, H. Huang, and Y. Xia, "H2NF-Net for brain tumor segmentation using multimodal MR imaging: 2nd place solution to BraTS challenge 2020 segmentation task," 2020, *arXiv:2012.15318*. [Online]. Available: <http://arxiv.org/abs/2012.15318>
- [48] T. Henry, A. Carre, M. Lerousseau, T. Estienne, C. Robert, N. Paragios, and E. Deutsch, "Brain tumor segmentation with self-ensembled, deeply-supervised 3D U-Net neural networks: A BraTS 2020 challenge solution," 2020, *arXiv:2011.01045*. [Online]. Available: <http://arxiv.org/abs/2011.01045>
- [49] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Feb. 2021.
- [50] S. Bourouis, I. Channoufi, R. Alroobaea, S. Rubaiee, M. Andejany, and N. Bouguila, "Color object segmentation and tracking using flexible statistical model and level-set," *Multimedia Tools Appl.*, vol. 80, no. 4, pp. 5809–5831, 2020.
- [51] O. Maier, B. H. Menze, J. von der Gabelntz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, and L. Chen, "ISLES 2015—A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI," *Med. Image Anal.*, vol. 35, pp. 250–269, Jan. 2017.
- [52] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, and P. Büchler, "The virtual skeleton database: An open access repository for biomedical research and collaboration," *J. Med. Internet Res.*, vol. 15, no. 11, p. e245, Nov. 2013.
- [53] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, 2015.
- [54] R. Alroobaea, S. Rubaiee, S. Bourouis, N. Bouguila, and A. Alsufyani, "Bayesian inference framework for bounded generalized Gaussian-based mixture model and its application to biomedical images classification," *Int. J. Imag. Syst. Technol.*, vol. 30, no. 1, pp. 18–30, 2020.



ROOBAEA ALROOBAEA received the bachelor's degree (Hons.) in computer science from King Abdulaziz University (KAU), Saudi Arabia, in 2008, and the master's degree in information systems and the Ph.D. degree in computer science from the University of East Anglia, U.K., in 2012 and 2016, respectively. He is currently an Associate Professor with the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include human-computer interaction, software engineering, cloud computing, the Internet of Things, artificial intelligence, and machine learning.



SAQIB QAMAR received the master's degree from Aligarh Muslim University and the Ph.D. degree from the Huazhong University of Science and Technology, China. He is currently working as an Assistant Professor with the Madanapalle Institute of Technology and Science (MITS). His research interests include computer vision, medical image analysis, and deep learning.



RAN ZHENG received the M.S. and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), China, in 2002 and 2006, respectively. She is currently an Associate Professor of computer science and engineering at HUST. Her research interests include distributed computing, cloud computing, and high-performance computing.



FADY ALNAJJAR received the M.S. degree in artificial intelligence and the Ph.D. degree in system design engineering from the University of Fukui, Japan, in 2007 and 2010, respectively. Since 2010, he has been a Research Scientist with the Brain Science Institute (BSI), RIKEN, Japan. He conducted a neuro-robotics study to understand the underlying mechanisms for embodied cognition and the mind. In 2012, he started interested in exploring the neural mechanisms of motor learning, adaptation, and recovery after brain injury from the sensory- and muscle-synergies perspectives. His research target is to propose an advanced neuro-rehabilitation application for patients with brain injuries.



FATHIA ABOUDI is currently pursuing the Ph.D. degree in biophysics and medical imaging with the High Institute of Medical Technology in Tunisia (ISTMT). She is a Research Member with the Research Laboratory of Biophysics and Medical Technology (LRBTM). In 2019, she was registered in Complementary Studies Certificates (CEC) Post University Degrees in MRI in clinical practice at the Faculty of Medicine in Tunisia (FMT). She has experience in medical imaging and nuclear medicine. She is interested in the combination between the medical field and high-tech.



PARVEZ AHMAD received the B.S. and M.S. degrees from Aligarh Muslim University, India, in 2009 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Huazhong University of Science and Technology (HUST), China. His research interests include deep learning, high-performance computing, cloud computing, medical imaging, and computer vision.



HAI JIN (Fellow, IEEE) received the Ph.D. degree in computer engineering from the Huazhong University of Science and Technology (HUST), China, in 1994. He worked at The University of Hong Kong, from 1998 to 2000. He worked as a Visiting Scholar at the University of Southern California, from 1999 to 2000. He is currently a Chair Professor of computer science and engineering with HUST. He has coauthored 22 books and published over 800 research papers. His research interests include computer architecture, virtualization technology, cluster computing and cloud computing, peer-to-peer computing, network storage, and network security. He is a fellow of CCF and a member of ACM. In 1996, he was awarded a German Academic Exchange Service Fellowship to visit the Technical University of Chemnitz, Germany. He was awarded the Excellent Youth Award from the National Science Foundation of China, in 2001.