Hybrid-DANet: An Encoder-Decoder Based Hybrid Weights Alignment with Multi-Dilated Attention Network for Automatic Brain Tumor Segmentation

Naveed Ilyas, Yoonguu Song and Boreom Lee (*Member IEEE*)

***Abstract*—Gliomas are the most common and highly growing tumors lead to high mortality rate in their highest grade. The early diagnosis of gliomas, and treatment planning are most important steps to enhance the life expectancy of a patient. Among the modern imaging techniques, magnetic resonance imaging (MRI) is the most robust and widely used technique to visualize the brain tumor. The CNN-based networks mainly depend on multi-branch and increasing the depth/width of the network to enhance the segmentation accuracy at the cost of high computational cost. To mitigate these drawbacks we therefore, propose a hybrid weights alignment with multi-dilated attention network for automatic brain tumor segmentation (Hybrid-DANet). It employs multiple modules incorporated on baseline encoder-decoder architecture. Firstly, we proposed a novel hybrid weight alignment with multi-dilated attention module (HWADA) is used between the skip connections. It has capability to obtain the different sets of aligned weight by using different dilation schemes. Different weight alignments play a vital role to obtain very precise targeted information while negating the less informative part. It utilizes the low and high level information with skip connections across each branch of encoder and decoder. Secondly, we incorporated a multi channel multi scale module (MCS) on the baseline module. It consists of multiple channels used to extract the channel-wise information with more reduced computational cost. To reduce the resultant saturated accuracy due to vanishing gradient problem, we incorporated the residual module (RM). Thus, the RM, and MCS are useful to obtain the deep, intrinsic, channel-wise feature without expansion of depth and height. Whereas the novel HWADA not only propagates the low level information but also process it to obtain more semantic features used for decoder. We have tested our proposed technique on well-known datasets; BraTS 2017, and 2018 with comparable performance to counter-part.**

***Index Terms*—Deep learning, Medical Image Analysis, Brain Tumor.**

I. Introduction

BRAIN is one of the most important and sensitive part of human body. The most prevalent with high mortality rate tumors are called gilomas [1]. It is further graded into low grade gilomas (LGG) and high grade gilomas (HGG), with the latter being more aggressive and high growth rate than the former. The brain tumor treatments include radiotherapy, surgery, and chemotherapy. The early diagnosis of brain tumor plays a significant role to reduce the mortality rate. The most complementary and prominent information regarding tumor is obtained through Magnetic Resonance Imaging (MRI) which employs four modalities: T1-weighted, T2-weighted, T1c, and Flair. These modalities further provide the robust information about the types of tumor. The more precise segmentation of tumor structure plays a vital role for treatment planning, accurate surgery, and follow up analysis. As manual segmentation of brain tumor is prone to error and time consuming, thus scientists are trying to develop autonomous system for brain tumor segmentation. However, the shape, structure, and location variations are most common challenges for accurate segmentation. Also, the neighbouring tissues arrangement is highly effected due to existence of tumor. The ability of automatic feature extraction in computer vision is improved by recent development of convolutional neural network (CNN). Due to this advancement, researchers are trying to propose state of the art methods for the precise and robust segmentation of brain tumor. Generally U-Net [20] and Fully Convolutional Network (FCN) [21] are most common and reliable methods in medical image segmentation. Among them, earlier one has shown most promising results. The U-shaped architecture employs encoder and decoder with skip connections are used to obtain low to high level features information.

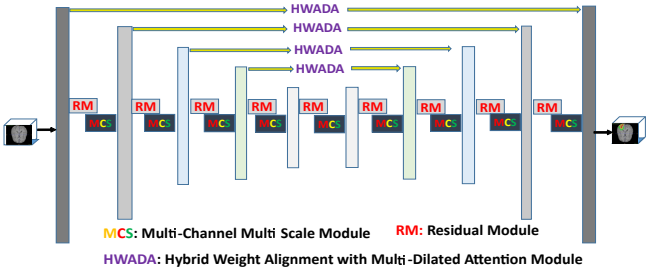

Figure 1. An Overview of Hybrid-DANet: The proposed network is composed of baseline U-shaped encoder-decoder module with integration of 1) Multi-channel multi-scale module (MCS), 2) Residual module (RM), 3) and hybrid weight alignment with dilated attention module (HWADA).

The structure of brain and their visual appearance are two main indicators to analyze the different types of tumors. In the recent past, a number of CNN-based algorithms have bee proposed to mitigate the challenges of brain tumor segmentation. In CNN, U-Net based architecture are prevalent and thus provide the better results in medical image segmentation. Firstly, existing U-Net based architectures enhance the segmentation accuracy by using multi-column architecture. For example, authors in [1] and [2] used multi-column architecture by taking advantages of different receptive fields. The specific set of images are targeted due to fixed kernel size in each column. Further, multi-column architecture used in [1] and [2] failed to obtain the high segmentation accuracy due extracting similar type of features [3]. Secondly, the network width and depth is increased to obtain the high accuracy irrespective

of their computational cost. IN other words, the CNN-based method are expanded depth or width-wise irrespective of their enhanced learning parameters. Thirdly, existing U-Net based architecture mostly import the existing modules without focusing on their own contribution to a specific module. For example, authors in [4], [5], and [6] integrated the existing module in the U-Net architecture to obtain the state of the art accuracy.

Based on these observations, we proposed a Hybrid-DANet: An encoder-decoder based hybrid weights alignment with multi-dilated attention network for Automatic Brain Tumor Segmentation. Our model comprises of four types of network, (i) encoder-decoder baseline module, (ii) residual module (RM) (iii) multi-channel multi-scale module (MCS), and (iv) a novel hybrid weights alignment with multi-dilated attention module (HWADA). Encoder-decoder is very powerful for baseline model for brain tumor segmentation to obtain the coarse to fine level features. MCS is very useful to obtain the channel-wise features with reduced computational complexity. Also channel wise feature extraction at multiple stages with different filter sizes enhances the final segmentation accuracy. HWADA is useful to extract the rich contextual targeted information in a specific scene. The extraction of targeted features capability is further enhanced by inclusion of contextual information through multi- dilation rates. The HWADA is applied across each skip connection of encoder-decoder architecture which is further useful to propagate the low level features to high level layers at the decoder side.

In summary, the main contribution of our research are as follows:

- We design a hybrid weights alignment with multi-dilated attention network to obtain abrupt to continuous varying scale features. The baseline encoder-decoder architecture with inclusion of RM and MCS enhance the low to high level feature extraction capability by eliminating the vanishing gradient problem. Also the channels-wise features are extracted with reduced computational cost.
- A novel hybrid weights alignment with multi-dilated attention module (HWADA) is proposed to obtain two types of different weight alignments, useful to obtain contextual target information while negating the less informative scene.
- Combination of MCS with HWADA incline the algorithm to obtain the varying geometrical information of each tumor type. MCS with multi-channel multi-scale approach is capable to exploit the channel inter-dependencies while reducing the number of parameters. Whereas HWADA with varying dilation rates obtain the targeted features with the help of contextual information.
- Extensive experiments are conducted on two challenging datasets depicted that our model achieves the state of the art performance.

II. Related Work

With the rapid growth of CNN-based techniques in classification, recognition, and especially segmentation tasks, the CNN-based methods are employed for the purpose of medical image analysis. To address challenges such as structure-variation and small amount of training samples, a number of researchers have contributed to enhance the segmentation accuracy.

Authors in [7] proposed CNN based brain tumor segmentation algorithm. The CNN is trained to learn the mapping from MRI space to tumor marker space. The predicted labels obtained from CNN are feed into support vector machine (SVM) for classification of different types of tumors. Authors in [8] proposed an automatic brain tumor segmentation using deep learning. The proposed network exploits both local and global contextual features simultaneously. Also, two phase training procedure is utilized to resolve the class imbalance problem. An automatic segmentation method is proposed by authors in [9] . They used the filter size of 3×3 with deeper architecture. Authors in [1] proposed a multi-scale CNN for brain tumor classification and segmentation.The input image is directed to three different spatial scale along different processing pathways. Authors in [4] proposed a U-Net based model with encoding and decoding structure, a residual module, and a spatial dilated feature pyramid (DFP) module, namely, DFP-ResUNet. A multiple convoluational layers with different dilation rates are used in parallel to obtain multi-scale features. Authors in [10] proposed UNet based fully convolutional networks for brain tumor segmentation. After applying CNN, extremely randamized trees classifier is used to segment the different classes of tumor. Authors in [11] proposed a novel idea to segment three different types of tumors (whole tumor, tumor core, enhancing tumor). They decomposed the multi-class segmentation into three binary segmentation problem. Firstly, the whole tumor is segmented, and the bounding box of result is utilized for tumor core segmentation. Similarly, the enhancing tumor is segmented on the bounding box of tumor core. Inspired from multi-column architecture of CNN, authors in [2] proposed hybrid two track U-Net architecture for brain tumor segmentation. The two tracks have different number of layers and filter size used to extract multiple size features. Finally, these two tracks are merged to obtain the final feature map. Authors in [5] presented an integrated architecture, combined different existing modules with UNet to obtain the final feature map. The UNet is used as a baseline architecture with incorporation of residual model and attention gate. The attention gates are used in skip connection to obtain the targeted information, propagated to decoder. Author in [6] is proposed modified U-Net architecture by including the residual extended skip and wide context module to obtain the low level features with contextual information for final feature map.

III. Proposed Model

The architecture of the proposed approach is shown in Fig. 2. Firstly, brain tumor segmentation start with concatenation of the four modalities into a single three dimensional volume, later feed to convolutional neural network (CNN). Secondly, Hybrid-DANet is built on encoder-decoder architecture (baseline). The proposed model employs deep multi-channel multi-scale module (MCS), residual module (RM), and hybrid

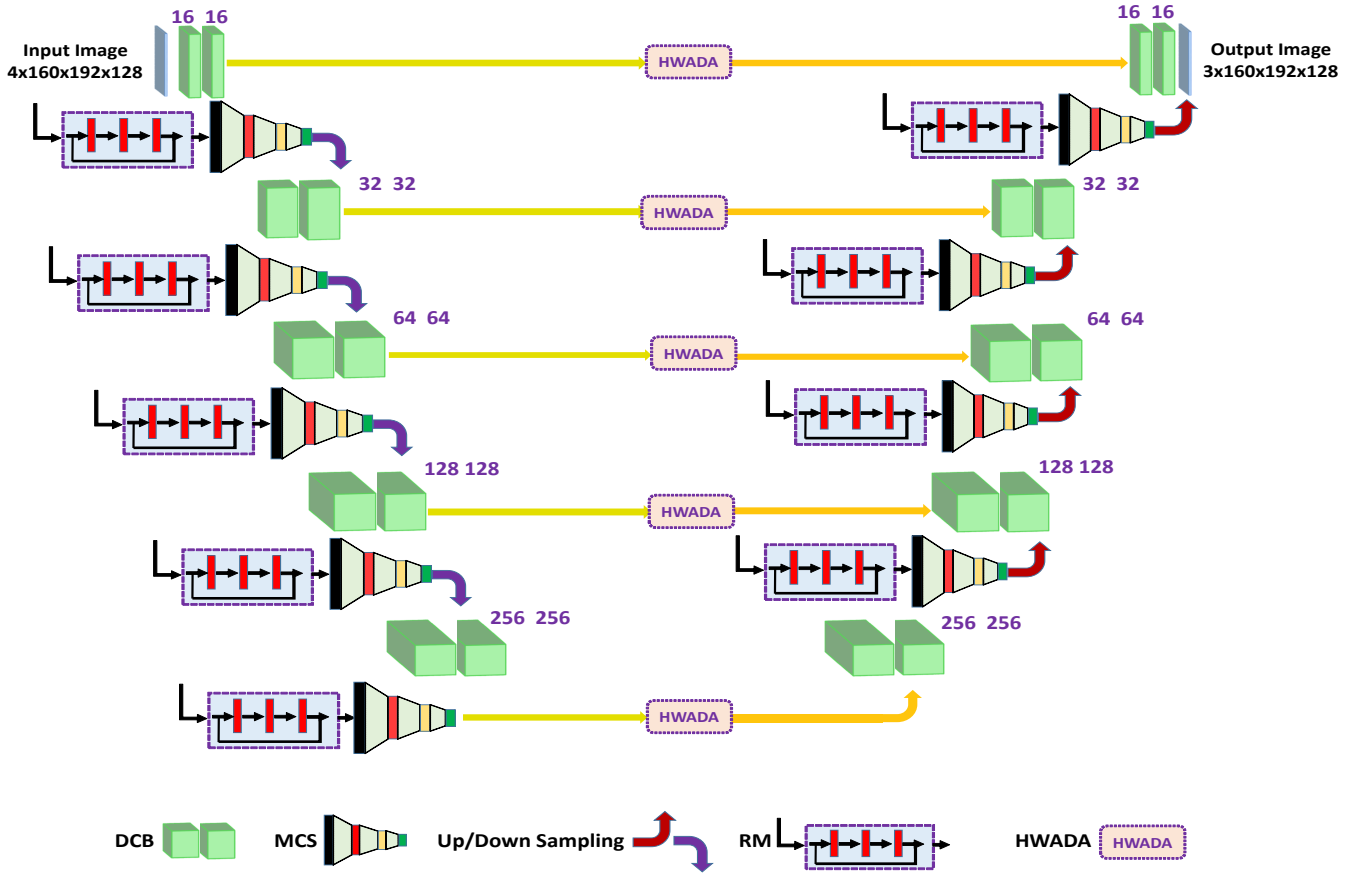

Figure 2. Hybrid-DANet: An encoder-decoder based hybrid weights alignment with multi-dilated attention network for automatic brain tumor segmentation. A baseline U-shaped encoder-decoder architecture employs double convolutional block (DCB) with integration of three sub-modules: 1) residual module (RM) ,2) multi-channel multi-scale module (MCS), and 3) hybrid weight alignment with dilated attention module (HWADA). The input MRI image first enter into DCB and is forwarded to RM, and MCS, respectively before down-sampling. This process is repeated in the encoder part, whereas, the same process is repeated with addition of HWADA through skip connections at the decoder, respectively.

weight alignment with dilated attention module (HWADA) incorporated on baseline. Thirdly, the segmentation of different types of tumors is quite difficult due to varying geometry (shape and structure) of tumors. Therefore, combination of RM, MCS and HWADA with baseline result in the extraction of robust, deep, multi-scale, and targeted information about each tumor type.

A. Hybrid-DANet: An encoder-decoder based dilated attention residual multi-scale network for automatic brain tumor segmentation

1) Baseline Model: It is a U-shaped encoder-decoder architecture with multiple double convolutional blocks (DCBs) are embedded in both encoder and decoder. The DCB consists of two convolutional layers with filter size of $3 \times 3 \times 3$ with group normalization (GN) and ReLU are placed at alternative position as shown in Fig. 6. There are total 10 DCB, 5 on each side. The five DCBs are placed on different levels on left leg of U-shaped architecture. In each level the number of channels are getting doubled whereas the spatial dimensions are halved on each level of encoder. Similarly, the five DCBs are used at five different levels of decoder. The number of channels are getting halved in each level whereas the up-convolution is used to double the spatial dimension of input MRI image. Whereas, the up-convolution consists of

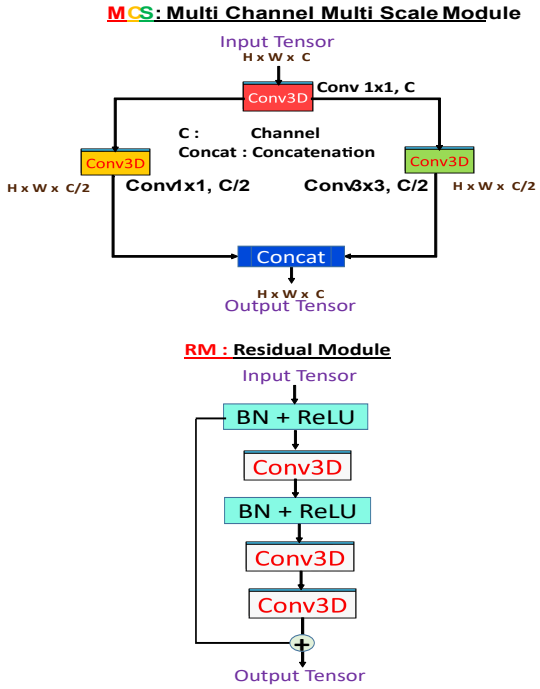

Figure 3. Hybrid-DANet: An encoder-decoder based hybrid weights alignment with multi-dilated attention network for automatic brain tumor segmentation. 1) a multi-channel multi-scale module (MCS) (top) and 2) a residual module (RM) (bottom).

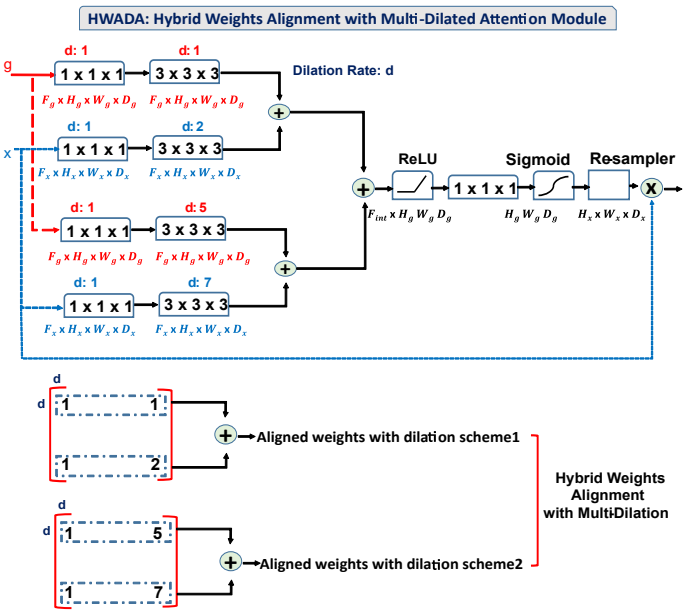

Figure 4. An overview of HWADA: A hybrid weights alignment with multi-dilated attention module. 1) Detailed description of HWADA (top). The expansion of HWADA i) aligned weights with dilation scheme 1 ii) aligned weights with dilation scheme 2 (bottom).

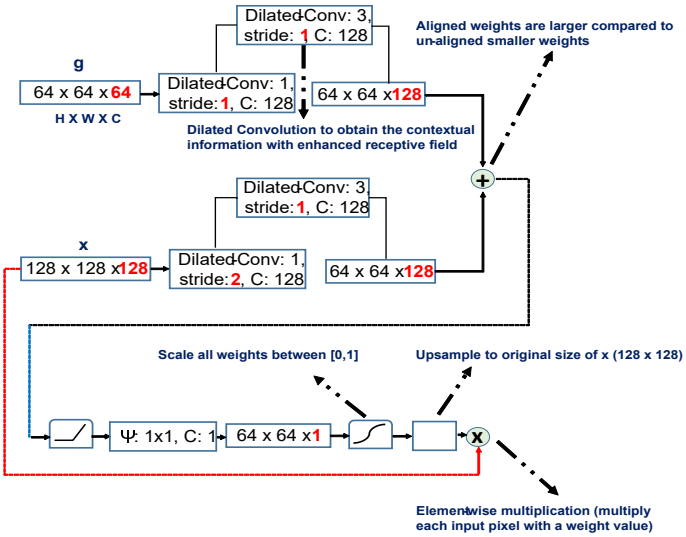

Figure 5. An expansion of HWADA (aligned weights with dilation scheme 1). The detailed description of each sub module is depicted through arrows.

up-sampling (factor:2), a convolution layer, GN, and ReLU are sequentially placed. Starting with an input MRI image of dimension $4 \times 160 \times 192 \times 128$ given to first DCB. It will sequentially moved to next lower level with reduction in spatial dimension while doubling number of channels. Also skip connections are applied after each levels to propagate the information from encoder to decoder. The deeper features are extracted at the encoder by increasing the number of channels while reducing the spatial dimension. Whereas, the spatial features are obtained in each level of decoder as dimensions are getting double. In addition, low level information are propagated via skip connections in each level to further enrich the features for better segmentation accuracy. As a summary, it is a very powerful baseline architecture with numerous features (discussed above) useful for brain tumor segmentation.

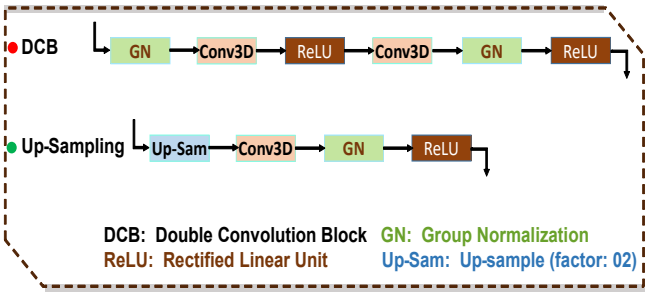

Figure 6. A detailed description of double convolutional block (DCB) at the top, whereas Up-sampling is expanded at the bottom. The DCB consists of GN, convolution (3 x 3 × 3) and RELU. And Up-sampling comprises of GN, convolution (3 × 3 × 3), ReLU, and up-sampling factor equal to 2, respectively.

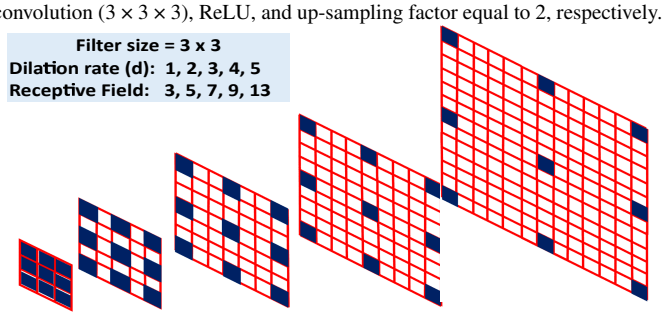

Figure 7. Dilated convolution with filter size 3×3 and increasing dilation rate of 1, 2, 3, 4, 5. Also visualization of receptive fields 3, 5, 7, 9, 13 are shown as well.

2) Multi-Chanel Multi-Scale Module (MCS): The structural variation, reduced training data, and image quality are the common challenges faced during brain tumor segmentation. To mitigate these challenges we propose a multi-channel, multi-scale module named as MCS inspired from [12]. The proposed module comprises of one stages with different filter size and channels as shown in Fig.3 (top). The module is capable to perform two types of tasks, (i) extraction of deep features using multiple scales and, (ii) channel-wise extraction of features. In the first step, the input tensor is squeezed by using $1 \times 1 \times 1$ convolution. Whereas, the channels are reduced by factor of two with different filter sizes of $1 \times 1 \times 1$ and $3 \times 3 \times 3$. The multiple size of filters are useful to obtain small to large scale information. As the size of tumors are varying thus by using different size of filters enhance the robustness of algorithm. Secondly channel-wise information is also useful to obtain the deep features along with spatial features obtained by multi-scaling. Also computational cost is minimized with reduced number of parameters. In this way, the MCS is capable to obtain spatial to deep features while reducing the size of parameters. As we have used the MCS in each stage of encoder-decoder architecture, the overall computational cost is reduced.

3) Residual Module (RM): Deeper network normally useful to obtain the deep rich features, however with increase number of layers result in saturated accuracy due to vanishing gradient problem. To avoid performance degradation due to vanishing gradient problem, we introduce the residual module. He et al.[13] proposed a residual learning correction scheme to avoid performance degradation which is expressed in Eq. 1

$$y = F(x, W_i) + x \tag{1}$$

Here x and y are the input and output tensors of the residual block, whereas W_i is the weight of i_{th} layer. $F(x, W_i)$ is the residual function added to input tensor x. By taking the partial derivative of y w.r.t x, the vanishing gradient problem is solved as the gradient does not disappear with the increase in number of layers.

The training accuracy increases as the number of layers increases, however the validation accuracy is badly disturbed due to vanishing/exploiting gradient problem. Thus the depth of the network is not directly proportional to the accuracy of an algorithm. Deep neural network (DNN) are difficult to train due to vanishing gradient problem. Motivated by the author in [13] and [14], we incorporated the residual module (RM) in the proposed network. The total 8 RM are embedded in both encoder and decoder. The structure of RM is very simple starting with batch normalization (BN) plus ReLU (non-linearity). The output of non-linear function is passed to 3D convolution layers followed by (BN + ReLU) as shown in Fig. 3 (bottom). Finally the output is passed to the two 3D convolution layers. The skip connection are applied to take the activation from lower layer and feed to higher layer. Due to strong feature learning property of RM, we incorporated 4 RM to encoder side followed by 4 RM to the decoder side. This will help to not only extract the low, medium, and high level features but also propagated to next layers using skip connections.

4) Hybrid Weight Alignment with Multi-Dilated Attention Module (HWADA): Contextual information aggregation, and better quality of output feature maps are achieved by using dilated convolution. It can be define as exponential increment of receptive field by intact the network parameters as depicted in Fig. 7. CNN with dilated convolution have proven to provide better performance in image segmentation. Therefore, by incorporating the dilated convolution in attention module, we greatly increase the ability of the network to selectively aggregate multi-scale contextual information. By observing and analyzing the benefits of dilated convolution, we incorporated dilated convolution in the baseline architecture with two variants as shown in 4 (bottom):

- Aligned weights with dilation scheme 1
- Aligned weights with dilation scheme 2

Both dilation rate variants are work in parallel as shown in Fig. 4 (top). As the attention gate (AG) are very popular to highlight the targeted area. This task is achieved by weight alignment in the existing AG. Another advantage of using AG is to obtain the rich targeted features transferred to decoder for better segmentation accuracy. By analyzing the existing advantages of AG, we realized that there is a room of improvement in the AG by incorporating the different sets of contextual information which results in different sets of weight alignments lately added to obtain a feature map with targeted contextual information which results in better segmentation accuracy. We therefore, introduce the two different ways of weights alignment by using multi-dilation schemes. The first scheme has dilation rate of 1, 1 (first row), 1, 2 (second row). The combination of these dilation rates are useful to obtain one type of contextual information (type 1 contextual information) by providing an output of aligned weights (type 1 aligned weights) Whereas, the second scheme has dilation rate of 1, 5 (first row), and 1, 7 (second row). The combination of these dilation rates are useful to obtain second type of contextual information (type 2 contextual information) by providing an output of aligned weights (type 2 aligned weights). These sets of aligned weights (type 1 aligned weights, and type 2 aligned weights) are added to obtain the feature map with rich targeted contextual information.

The more descriptive and detailed information of HWADA (aligned weights with dilation scheme 1)is depicted in Fig. 5. It has two inputs g and x, whereas g is coming from lower layer at the decoder side, whereas x is coming from encoder. These two inputs are passed through their respective branches of HWADA with different dilatation rates which results in different contextual information. And finally these different contextual information provide different sets of aligned weights. Due to rich contextual targeted information in the aligned weights, the output features plays a vital role in the final segmentation accuracy when added to decoder.

In this way compared to existing AG, we have not only obtain the targeted contextual information, but also introduce the different ways of obtaining contextual information with different weights alignment. This results in accruing the better targeted information due to hybrid contextual information. In addition, compared to AG, more rich features are propagated to decoder through skip connections.

5) Information Aggregation and Propagation to Decoder: Traditionally encoder-decoder architecture are used to translate the natural languages. As the low level information extracted by encoder is propagated to decoder to combine the low and high level features to obtain better performance. Instead of obtaining the whole information, a precise and accurate information could be useful to enhance the overall performance of algorithm. To obtain the targeted information authors in [15] proposed an attention-based network to extract, and preserve the targeted information in natural language processing. Inspired by authors in [15], we therefore proposed a hybrid weight alignment with multi-dilated attention module (HWADA) to obtain the more precise and specific information. The HWADA is used in skip connection in the proposed model. It is multi-column structure with varying filter sizes and dilation rate as well. In the first column we used the filter size of $1 \times 1 \times 1$ with same dilation rate of 1 in all four rows. Whereas, the second column employs the filter size of $3 \times 3 \times 3$ with dilation rate 1, 2, 5, 7. The output of first two rows are concatenated whereas the output of third and fourth row are combined and finally these two outputs are combined and pass to non-linear activation function (ReLU). Next, filter size of $1 \times 1 \times 1$ is used with sigmoid to obtain the final output. We have used five HWADA in the proposed model. The HWADAs are responsible to progressively suppress the irrelevant features responses in the background while highlighting the rich contextual targeted information. The five HWADAs are used in skip connections by highlighting the salient features. Also multi-dilation rates are applied across the second column of HWADAs, thus useful to obtain the contextual information by increasing the receptive field size. Information extracted from coarse scale is used in gating to disambiguate irrelevant and

noisy responses in skip connections. As a result, HWADA is responsible to suppress the irrelevant information while highlighting the targeted features. In addition, it is beneficial to obtain the increased receptive field, aggregating the contextual information, and the enhancement of feature map resolution. Gradients originating from background regions are down weighted during the backward pass. This allows model parameters in shallower layers to be updated mostly based on spatial regions that are relevant to a given task.

IV. Experiments

In this section, we describe the whole experiment details starting from network architecture to evaluation of proposed method. Moreover, this section is further divided into three sub-sections: implementation details, comparison with state of the art, and architecture ablation. In addition, we explain the performance comparison of the proposed Hybrid-DANet on two well-known datasets.

A. Implementation Details

1) Datasets: To evaluate the proposed method, we used the publically available datasets BraTS 2017 and BraTS 2018 [16][17]. These datasets are released by the Multimodal Brain Tumor Segmentation Challenge (BraTS) that run in conjunction with the International Conference On Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2017 and 2018. The datasets are divided into training and validation sets with each sample has four modalities, i.e., fluid-attenuated inversion recovery (FLAIR), T1 weighting (T1), T1 weighted contrast enhancement (T1ce), and T2 weighting (T2). The MRI images are skull-stripped and re-sampled to an isotropic $1 \times 1 \times 1\ mm^3$ with image dimension of $240 \times 240 \times 240$. The ground truth of each MRI image with same dimension are manually labelled by the experts provided the manual segmentation results. There are three types of labels in the ground truth MRI image named as enhanced tumor (labelled as 4), the peritumoral edema (labelled as 2), and non-enhancing tumor (labelled as 1). By considering these labels, whole tumor combined areas of labels 1, 2, 4, whereas tumor core combined areas of labels 1,4. And lastly, the enhancing tumor are labelled as 4.

2) Network Configuration: The network configuration of Hybrid-DANet is shown in Table I. The proposed network is comprised of four main modules: baseline, RM, MCS, and HWADA. For the sake of understanding we only demonstrate the encoder part of whole architecture. The whole encoder part (left leg of Hybrid-DANet) is divided into five levels. Whereas, each level consists of three sub-modules; DCB, RM, and MCS. And the decoder part consists of four levels (each consists of DCB, RM, HWADA, and MCS). The HWADA is connected via skip connections in each level of the decoder. The spatial dimensions are decreasing by half in each level of encoder, whereas the channels are getting doubled in the subsequent levels. Whereas, the spatial dimensions are up-sampled in each level of decoder with reduction of number of channels. In this way the encoder plays a vital role to obtain deep, intrinsic features, whereas the spatial features are obtained at the decoder side. Also the HWADA on the skip connections are applied to further enhance the quality of features for the final prediction.

3) Training Details: The BraTS exhibits a severe class imbalance problem. The distribution of sub-classes illustrated in Table IV. The distribution of classes are severely imbalance, approximately 98.46% of voxels belong to the healthy tissue thus labeled as background. However, edema and enhancing tumor only cover 1.02% and 0.29% voxels of the whole data, respectively. Lastly, the lowest volume is covered by non-enhancing tumor with rate of only 0.23%. The prepossessing of data elevates the class imbalance problem up to some extent, but it still effects the brain tumor segmentation accuracy. To mitigate this class imbalance problem, we employ a combined loss function that integrates focal traversky loss (FTL) and generalized dice loss (GDL) as shown in Eq. 2.

$$Loss = L_{FTL} + L_{GDL} \tag{2}$$

where L_{FTL} and L_{GDL} respectively represent the focal traversky loss and generalized dice loss, which are correspondingly defined as Equations 4 and 5, respectively.

Tversky index (TI) is a generalization of Dice coefficient. It adds weight to false positive (FP) and false negative (FN) by using β coefficient as shown in Eq. 3.

$$TI(p, \hat{p}) = \frac{p\hat{p}}{p\hat{p} + \beta(1-p)\hat{p} + (1-\beta)p(1-\hat{p})} \tag{3}$$

If $\beta = \frac{1}{2}$, it is similar to regular Dice coefficient. The focal tversky loss L_{FTL} mainly focuses on hard examples by down weighting the easy ones. The L_{FTL} learn the hard examples by using γ coefficient.

$$L_{FTL} = \sum (1 - TI_c)^{\gamma} \tag{4}$$

$$L_{GDL} = 1 - 2\frac{\sum_i^L w_i g_i p_i}{\sum_i^L w_i (g_i + p_i)} \tag{5}$$

where c represents the class, L denotes the total number of labels, and w_i denotes the weight assigned to the ith label. Further, p_i and g_i represent the pixel value of segmented binary image and the binary ground truth image, respectively. We used PyTorch plateform [23] with NVIDIA GeForce GTX 3070 with 8GB memory.

4) Data Augmentation: During the training process, we cropped the input MRI image $240 \times 240 \times 155$ into multiple patches of size $192 \times 160 \times 128$. By cropping the MRI image the training data is increased. Whereas, the data augmentation is not performed for validation dataset.

B. Comparison with Existing Algorithms

1) Evaluation Metrics: We evaluated the Hybrid-DANet on two well known datasets (BraTS 2017, 2018). We utilized the most commonly used evaluation metrics such as dice similarity coefficient (DSC), sensitivity, specificity, and Hausdorff distance (HD) for the model evaluation. The DSC, sensitivity, specificity, and HD can be calculated by using Equations: 6, 7,8, 9, respectively.

Table I
THE NETWORK ARCHITECTURE OF HYBRID-DANET: AN ENCODER-DECODER BASED HYBRID WEIGHTS ALIGNMENT WITH MULTI-DILATED ATTENTION NETWORK FOR AUTOMATIC BRAIN TUMOR SEGMENTATION. WE HAVE USED FIVE COLOURS (RED, BLUE, GREEN, CYAN, BLACK) TO DIFFERENTIATE BETWEEN FIVE DIFFERENT LEVELS OF ENCODER.

Modules	Channels	Kernels	Padding	Dilation	Output Size	Hybrid-DANet
DCB1	16	3x3	1	1	16x128x128x128	Conv-16
RM1	16	3x3	1	1	16x128x128x128	Conv-16
MCS1	16,8,8,8,4	1x1, 3x3	1	1	16x128x128x128	Conv-16 Down-Sample
DCB2	32	3x3	1	1	32x64x64x64	Conv-32 Conv-32
RM2	32	3x3	1	1	16x128x128x128	Conv-32
MCS2	32,16,16,16,8	1x1, 3x3	1	1	32x64x64x64	Conv-32 Down-Sample
DCB3	64	3x3	1	1	64x32x32x32	Conv-64 Conv64
RM3	64	3x3	1	1	16x128x128x128	Conv-64
MCS3	64,32,32,32,16	1x1, 3x3	1	1	64x32x32x32	Conv-64 Down-Sample
DCB4	128	1x1, 3x3	1	1	128x16x16x16	Conv-128 Conv-128
RM4	128	3x3	1	1	16x128x128x128	Conv-128
MCS4	128,64,64,64,32	1x1, 3x3	1	1	128x16x16x16	Conv-128 Down-Sample
DCB5	256	3x3	1	1	256x8x8x8	Conv-256 Conv-256
RM5	256	3x3	1	1	16x128x128x128	Conv-256
MCS5	256,128,128,128,64	1x1, 3x3	1	1	256x8x8x8	Conv-256

Table II
THE QUANTITATIVE RESULTS OF HYBRID-DANET ON BRATS-2017 VALIDATION DATASET (57 MRI SCANS)

Method	Dice			Sensitivity			Specificity			Hausdorff		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
UNet[18]	0.751	0.564	0.147	0.775	0.650	0.1147	0.995	0.996	0.999	26.80	22.59	**14.65**
Atten-UNet[19]	0.509	0.396	0.339	0.591	0.503	0.464	0.982	0.984	0.986	**16.61**	21.25	20.74
SegNet[20]	0.833	0.703	0.496	-	-	-	-	-	-	-	-	-
PSPNet[21]	0.809	0.701	0.554	-	-	-	-	-	-	-	-	-
NovelNet[22]	0.876	0.763	0.642	-	-	-	-	-	-	-	-	-
Hybrid-DANet-v1	0.873	0.735	0.658	0.877	0.757	0.660	0.997	0.998	0.999	19.63	**19.85**	16.13
Hybrid-DANet	**0.892**	**0.761**	**0.680**	**0.883**	**0.757**	**0.704**	**0.998**	**0.999**	**0.999**	26.70	19.94	17.41

Table III
THE QUANTITATIVE RESULTS OF HYBRID-DANET ON BRATS-2018 DATASET (VALIDATION DATASET)

Method	Dice			Sensitivity			Specificity			Hausdorff		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
Baseline	0.8221	0.6764	0.5018	0.8289	0.6835	0.4938	0.9984	0.9982	0.9992	42.359	22.302	27.990
Baseline-Double-HWADA	0.8718	0.6834	0.5701	0.8797	0.7186	0.5985	0.9958	0.9964	0.9982	17.117	16.885	17.390
Baseline Multi-Scale Atten	0.8840	0.7620	0.6157	0.88828	0.7775	0.6231	0.9978	0.9977	0.9993	25.774	19.357	17.525
Hybrid-DANet-v1	0.8818	0.7118	0.6233	0.8725	0.7187	0.6252	0.9998	0.9983	0.9992	26.592	16.638	15.657
Hybrid-DANet	0.8771	0.7596	**0.6498**	**0.8901**	0.7362	**0.6573**	0.9983	**0.9994**	**0.9995**	22.1930	19.307	**14.590**

Table IV
THE DISTRIBUTION OF SUB-CLASSES IN BRATS DATASET.

Class	Rate %
Background	98.46
Edema	1.02
Enhancing Tumor	0.29
Necrotic and non-enhancing tumor	0.23

$$DSC = \frac{2TP}{FN + FP + 2TP} \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$HD = max(sup_{a \in A} inf_{b \in B} d(a,b), sup_{b \in B} inf_{a \in A} d(a,b) \tag{9}$$

TP, FP, FN stands for true positive, false positive, and false negative. Whereas HD is used to computes Hausdorff distance between the the binary objects in two images, sup is supremum, inf is the infimum and d is the absolute distance value.

2) Experiment Results on the BraTS 2017 Dataset: This experiment is conducted on BraTS 2017 dataset. It has total number of 285 MRI images. We have used 228 (80%) of data for training, whereas rest of 57 (20%) images are used for validation purpose. We compared the proposed method with baseline UNet and also with recent state of the art methods including Atten-Unet, SegNet, PSP-Net, and NovelNet as

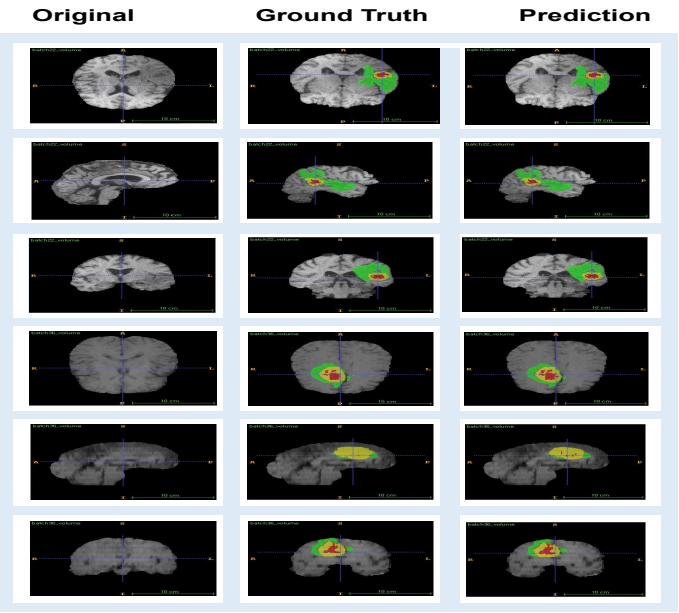

Figure 8. Visualization of BraTS-2017 dataset; MRI Image, ground truth, and prediction.

shown in Table II. The proposed method outperform the rest of the state of the art algorithms by achieving the mean dice score of 0.892, 0.764, 0.680 for the WT, TC, and ET, respectively.

3) Experiment Results on the BraTS 2018 Dataset: This experiment is conducted on BraTS 2018 dataset. To evaluate the proposed Hybrid-DANet, we have used 171 (75%) images for training, whereas, 57 (25%) images are used for validation purpose. We compared the proposed model with well known methods. We evaluate the proposed method by using four evaluation metrics (Dice, Sensitivity, Specificity, Hausdorff) on different types of tumor classes (WT, TC, ET). Table III depicts the performance of proposed method. A comparable performance is achieved on validation dataset. This is due to extraction of shallow to deeper features. The more intrinsic and deeper features are obtained through encoder whereas, the spatial features are combined with the target oriented features at the decoder side further enhance the accuracy on ET as shown in Table III.

C. Architecture Ablation

This subsection is devoted to understand the capability of each module in the proposed model. Also the placement of module in different places are analysed as well. The ablation study consist of five types of network.

- Baseline: It is an encoder-decoder based network.
- Baseline Double HWADA: It is an architecture with placement of HWADA on skip connections as well as on the encoder and decoder side after the first double convolution layers.
- Baseline Multi-Scale Attention: It is an architecture with MCS incorporated on the encoder-decoder side, whereas the attention modules are used with skip connections between encoder and decoder.
- Baseline Multi-Scale HWADA: It is an architecture with baseline encoder-decode network. Whereas, the MCS is incorporated in both encoder and decoder side. Further, the HWADA is used with skip connections as well.
- Hybrid-DANet: It is an architecture with inclusion of MCS, RM on the encoder and decoder side. And HWADA is used in each skip connections to further improve the segmentation accuracy of class imbalance problem.

We performed the ablation study on BraTS 2018 dataset. Four evaluation metrics (Dice, Sensitivity, Specificity, Hausdorff) are observed during the ablation study. Firstly, we start the evaluation with a well known encoder-decoder architecture named as UNet (baseline). It achieves the dice score of (0.8221, 0.6764, 0.5018) on WT, TC and ET, respectively. Secondly, we placed the attention module on skip connection and on the encoder-decoder side as well (Baseline-Double-HWADA). A reasonable performance increment is observed on WT and TC as shown in Table III. Thirdly, we used the MCS module on both encoder and decoder side with placement of attention module on the skip connection (Baseline-Multi-Scale-Atten). The performance is increased compared to Baseline-Double-HWADA. Fourthly, we used the HWADA on skip connections with MCS incorporated on both encoder and decoder side (Hybrid-DANet-v1). Lastly, RM, MCS are incorporated on encoder-decoder with HWADA is applied on skip connections (Hybrid-DANet). A reasonable performance increment is observed compared to rest of architectures on four evaluation metrics, Dice, Sensitivity, Specificity, and Hausdorff as shown in Table III. As we know the segmentation of ET is more difficult compared to WT, TC due to its small size (class imbalance problem). More specifically, the value of ET w.r.t the dice, sensitivity, Hausdorff, and specificity outperform the counterpart.

V. Conclusion and Future Work

In this work, we proposed a novel hybrid weights alignment network with multi-dilated attention to automate the brain tumor segmentation that is trained in an end-to-end manner. Due to strong feature extraction property of deep neural network, we used multi- scale residual attention networks to extract the deep and scale varying features. Furthermore, a HWADA is incorporated in baseline to obtained the targeted contextual information lately used for better feature extraction to enhance the segmentation accuracy. The performance of Hybrid-DANet is comparable to the state-the-art methods due to varying receptive field and strong ability of handling class imbalance issues. In this way, our proposed approach is capable of learning the low to complex, deeper and scale-aware, contextual, and targeted features with enhanced quality of feature map. In future, we intend to used the transformer in encoder-decoder architecture to further improve the overall accuracy.

References

[1] F. J. Díaz-Pernas, M. Martínez-Zarzuela, M. Antón-Rodríguez, and D. González-Ortega, "A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network," in *Healthcare*, vol. 9, p. 153, Multidisciplinary Digital Publishing Institute, 2021.

[2] N. M. Aboelenein, P. Songhao, A. Koubaa, A. Noor, and A. Afifi, "Httu-net: hybrid two track u-net for automatic brain tumor segmentation," *IEEE Access*, vol. 8, pp. 101406–101415, 2020.

[3] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1091–1100, 2018.

[4] J. Wang, J. Gao, J. Ren, Z. Luan, Z. Yu, Y. Zhao, and Y. Zhao, "Dfp-resunet: Convolutional neural network with a dilated convolutional feature pyramid for multimodal brain tumor segmentation," *Computer Methods and Programs in Biomedicine*, p. 106208, 2021.

[5] J. Zhang, Z. Jiang, J. Dong, Y. Hou, and B. Liu, "Attention gate resunet for automatic mri brain tumor segmentation," *IEEE Access*, vol. 8, pp. 58533–58545, 2020.

[6] M. U. Rehman, S. Cho, J. H. Kim, and K. T. Chong, "Bu-net: Brain tumor segmentation using modified u-net architecture," *Electronics*, vol. 9, no. 12, p. 2203, 2020.

[7] W. Wu, D. Li, J. Du, X. Gao, W. Gu, F. Zhao, X. Feng, and H. Yan, "An intelligent diagnosis method of brain mri tumor segmentation using deep convolutional neural network and svm algorithm," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.

[8] 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," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[9] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[10] H. T. Le and H. T.-T. Pham, "Brain tumour segmentation using u-net based fully convolutional networks and extremely randomized trees," *Vietnam Journal of Science, Technology and Engineering*, vol. 60, no. 3, pp. 19–25, 2018.

[11] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI brainlesion workshop*, pp. 178–190, Springer, 2017.

[12] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[16] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[17] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[19] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

[22] H. Li, A. Li, and M. Wang, "A novel end-to-end brain tumor segmentation method using improved fully convolutional networks," *Computers in biology and medicine*, vol. 108, pp. 150–160, 2019.

[23] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.